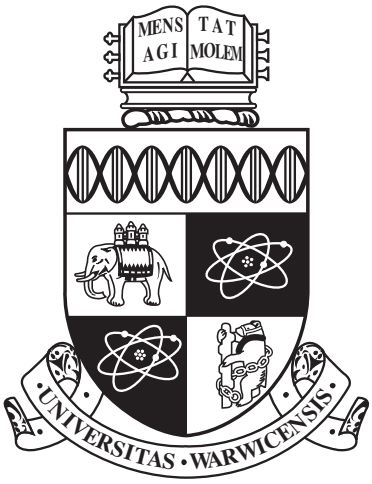

Finding Network Modules and Motifs Regulating Plant Stress Responses: Integration and Modelling across Multiple Data Sets

by

Krzysztof Polański

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Systems Biology

September 2015

THE UNIVERSITY OF
WARWICK

Contents

List of Tables **v**

List of Figures **vi**

Acknowledgments **viii**

Declarations **x**

Abstract **xi**

Abbreviations **xii**

Gene Names **xvii**

Chapter 1 Introduction **1**

1.1 Motivation 1

1.1.1 Feeding an exponentially growing population 1

1.1.2 *Arabidopsis thaliana* as a model organism in studying plant stress response 2

1.2 Stress response in *A. thaliana* 4

1.2.1 PAMP triggered immunity 6

1.2.2 Effector triggered susceptibility 7

1.2.3 Effector triggered immunity 9

1.2.4 Examples of plant pathogens 10

1.2.4.1 *Pseudomonas syringae* 10

1.2.4.2 *Botrytis cinerea* 11

1.2.5 Phytohormones in the plant stress response 13

1.2.5.1 Jasmonic acid 13

1.2.5.2 Salicylic acid 13

1.2.5.3 Abscisic acid 14

1.2.5.4 Ethylene 15
1.2.5.5 Crosstalk between the signalling of different phytohormones 16
1.3 Transcription regulation and gene regulatory networks 17
1.3.1 General overview of transcription 18
1.3.2 Searching for a network footprint 21
1.3.2.1 Clustering 22
1.3.2.2 Biclustering 23
1.3.3 Searching for gene regulatory network models 26
1.3.3.1 Undirected co-expression networks 26
1.3.3.2 Ordinary differential equation models 27
1.3.3.3 Mining large-scale transcriptomic datasets for gene regulatory networks 28
1.3.3.4 Experimentally derived networks 30
1.4 Aims and organisation of this thesis 32

Chapter 2 Wigwams: identifying gene modules co-regulated across multiple biological conditions 35
2.1 Introduction 40
2.2 Materials and methods 42
2.2.1 Input 42
2.2.2 Identifying modules spanning multiple datasets 43
2.2.3 Merging similar modules spanning the same time series subset 46
2.2.4 Sweeping redundant modules spanning different dataset subsets 49
2.2.5 GO term and TF binding motif enrichment testing 49
2.2.6 Yeast one-hybrid technique 49
2.3 Results 51
2.3.1 Wigwams systematically scans the data for evidence of co-regulation 51
2.3.2 Wigwams effectively removes redundancy among modules . . 52
2.3.3 Wigwams reveals expression signatures of regulatory mechanisms 53
2.3.4 Biological validation of detected modules 55
2.4 Discussion 58

Chapter 3 Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in

Arabidopsis leaves following infection with *Pseudomonas syringae* pv. *tomato* DC3000 **61**
3.1 Introduction . . . 65
3.2 Results . . . 68
3.2.1 Transcriptional dynamics of MTI and ETS revealed from a large-scale, highly-resolved time series expression study . . . 68
3.2.2 The majority of transcriptional changes are initiated by 6hpi 71
3.2.3 Early effector activity leads to major transcriptional changes prior to increased bacterial growth . . . 72
3.2.4 Detailed analysis of gene expression patterns during MTI and ETS . . . 77
3.2.5 Early sustained effector specific DEGs are predicted to modulate perception of external stimuli and chromatin re-organisation 81
3.2.6 Investigation of regulatory elements driving establishment of defence or disease . . . 82
3.2.7 Multiple time series co-expression analysis predicts specific regulation of pathogen-responsive genes . . . 85
3.2.8 Modelling the transcriptional network topology during disease and defence . . . 88
3.3 Discussion . . . 94
3.3.1 The chronology of effector-mediated transcriptional reprogramming . . . 94
3.3.2 Biological processes impacted by effectors . . . 96
3.3.3 An emerging role for chromatin remodeling early in the susceptible interaction . . . 97
3.3.4 Predictions of regulatory relationships underlying MTI and ETS 98
3.3.5 Network modelling highlights key effector-modulated genes . 99
3.4 Methods . . . 100

Chapter 4 Inference of functional gene regulatory networks mediating Arabidopsis response to environmental stress **105**
4.1 Introduction . . . 109
4.2 Results . . . 110
4.2.1 Elucidating TF network models underlying Arabidopsis responses to abiotic and biotic stress . . . 110
4.2.2 Identifying regulatory footprints in transcriptome data . . . 113
4.2.3 Expanding the transcription factor-only network models . . . 122

4.2.4 Functional inference for TFs mediating *B. cinerea* and *P. syringae* defence response 124
4.2.5 Identification of a combinatorial regulation network module predicted to play a role in hormone signalling 128
4.3 Discussion 131
4.4 Methodology 136
4.4.1 Inference of TF-only Network Models 136
4.4.2 Identifying Wigwams modules 137
4.4.3 Functional inference of Wigwams modules 137
4.4.4 Network expansion and functional analysis 137

Chapter 5 Discussion **139**
5.1 Significance of the work 139
5.2 Discussion of the applied methodology 141
5.3 Future work 146

List of Tables

2.1 Gene module information during Wigwams analysis 48

4.1 The distribution of Wigwams modules spanning each pair of conditions 115
4.2 The comparison of the number of network edges in the original transcription factor-only M-VBSSM models with the networks expanded with Wigwams modules 124
4.3 The number of TFs in the individual pathogen response networks, as well as the intersection, with predicted downstream functionality . . 127
4.4 The twelve selected genes and their prior (GO terms present in annotation) and inferred (result of functional analysis performed in the study) defence hormone related functionality 128
4.5 The number of shared downstream targets between the five genes forming a hormone signalling module in the intersection network . . 130

List of Figures

1.1 The zigzag model, capturing the interactions between the plant and the pathogen . . . 5
1.2 Crosstalk between regulatory networks in *A. thaliana* response to drought and cold . . . 19
1.3 Comparison of clustering and biclustering . . . 22

2.1 Strong evidence for dependent co-expression is detected during the module identification stage . . . 44
2.2 Merging . . . 47
2.3 Sweeping . . . 50
2.4 Four examples of modules showing different regulatory phenomena detected by Wigwams . . . 54
2.5 The number of modules identified for each combination of conditions 56
2.6 Wigwams modules are enriched in GO-terms and TF binding motifs 57

3.1 Dynamics of differentially expressed genes during basal defence and disease development . . . 70
3.2 Time at which gradients of DEGs begin to significantly differ between treatments . . . 73
3.3 Growth curves of DC3000 and DC3000*hrpA*-, with selected GO terms enriched by genes changing expression at indicated time points . . . 75
3.4 Response categories of DEGs capturing different MTI and ETS profiles and their validation . . . 78
3.5 Revealing links between TF binding motifs and temporal expression patterns . . . 83
3.6 Genes containing the same transcription factor binding site(s) in their upstream promoter sequences are co-expressed across multiple conditions . . . 87

3.7 The inferred transcription factor network model, jointly obtained for mock, DC3000*hrpA*- and DC3000 90
3.8 The expression profiles of three genes present in the inferred transcription factor network model 92

4.1 The process of obtaining M-VBSSM network models 112
4.2 The number of Wigwams modules identified for each condition combination and the distribution of module totals across the number of conditions they span 114
4.3 The overrepresentation of GO term groups among gene members of Wigwams modules 117
4.4 The overrepresentation of regulatory motifs bound by known transcription factors in the promoters of genes identified as part of Wigwams modules 119
4.5 The expression of two example Wigwams modules identified by analysing the six Arabidopsis time course datasets 120
4.6 The primary process of expanding M-VBSSM networks with non-transcription factor genes using Wigwams modules 123
4.7 The expanded M-VBSSM network connections that are in common to both the *B. cinerea* and *P. syringae* infection models and functional inference for expanded M-VBSSM networks 126
4.8 A truncated view of the five-gene combinatorial regulation hormone signalling interaction identified in the *B. cinerea*/*P. syringae* intersection network 129

Acknowledgments

Dr Katherine Denby, for being the model supervisor, giving me ample opportunities to collaborate on multiple projects and publications, and introducing the word gobbledygook to my vocabulary.

Dr Sascha Ott, for valuable feedback keeping me focused on the big picture and allowing me to hone my teaching skills at the programming course.

Dr Bärbel Finkenstädt, for keeping me in touch with the statistical realm, introducing me to the benefits of using splines for data processing, and recommending APTS courses - the Glasgow one was particularly enjoyable!

Prof. Jim Beynon, Dr Miriam Gifford, Dr Daniel Hebenstreit and Dr Simon Spencer, for constituting my advisory panel, offering helpful suggestions, and keeping me on track with the project.

The Warwick Systems Biology Doctoral Training Centre, for the funding which allowed me this opportunity — in spite of me taking a month to reply due to faulty email forwarding between accounts!

Prof. Murray Grant and Dr Laura Lewis, for letting me get involved in the *Pseudomonas syringae* analysis and opening my eyes to the process of biological interpretation by finding sense in my initial run results.

Dr Christopher Penfold, for his endless patience in explaining the workings of VBSSM, GP2S and CSI. Dr Dafyd Jenkins, for all-around aid, including Cytoscape expression visualisation and the gradient tool. All the members of the PRESTA project, for giving me multiple chances to work on my presenting skills and providing insightful comments on my research.

Christine Hicks, for being the first friend I made on arrival and always offering

calm support in the face of adversity. Claire Stoker, for being the perfect down to earth conversation partner, offhand knowledge of haustoria, and showing me leaf seven in the face of adversity. Lennie Foster, for educating me how not to accidentally burn the greenhouse down and monkey cookies. The rest of the C030 people, past and present, for a healthy dose of camaraderie both on and off campus. All of you folks are awesome and make me wish I had come out of my shell quicker.

José Ricardo Camões de Oliveira, for being the greatest friend imaginable, and turning out to exist in the flesh once we were both located in the UK. Also for making me look up how to make all those weird characters in LaTeX. Soon enough you'll be looking how to make a ń, mark my words!

Last, and certainly not least, my mom, dad and sister. I wouldn't be here if not for your never-ending support and warmth only a family can give, be it in the form of picking me up from the airport at ungodly hours, casual EM discussion during walks, or simple long-distance conversations about daily minutiae. I dedicate this thesis to all of you.

Declarations

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree except where otherwise stated. The work in this thesis has been undertaken by myself except where otherwise stated.

Abstract

In spite of constant technological advancements, world hunger remains a major challenge due to exponential population growth, and the loss of effectiveness of crop treatments such as pesticides. As such, comprehending the plant response to stress is of great importance in breeding more resilient crops. Whilst different stresses elicit distinct responses from the plant, a core set of regulatory interactions are conserved across multiple responses and operate as networks.

In this thesis, computational approaches were used to elucidate such regulatory interactions from time course expression datasets, predominantly through identification of genes co-expressed across multiple stimuli responses as a footprint of shared network co-regulation. The identification of such network footprints was tackled through Wigwams, a data mining algorithm capable of detecting groups of genes co-regulated across multiple datasets. In contrast to other algorithms, Wigwams assesses whether the co-expression it detects is likely to reflect co-regulation. The modules it found were significantly enriched in functionality and cis-regulatory elements, indicating actual co-regulation.

Wigwams and other computational approaches were applied to time course expression data capturing *Arabidopsis thaliana* response to *Pseudomonas syringae* pv. *tomato* DC3000. The presence of a virulent and avirulent strain in the experiment allowed for the temporal deconstruction of the regulatory events underlying the virulent strain's attempts to overcome plant defence through effector action. This analysis led to the detection of a number of effector-specific transcription changes stifling the defence response and manipulating the host's gene and protein expression. A transcription factor-only regulatory network model was proposed to explain the detected network footprints.

The inference of causal regulatory networks from expression data is a daunting task, and transcription factor-only models are a good computational compromise by capturing the key regulatory events taking place. However, they are lacking in target genes that carry out the functionality induced by the signalling, making functional assessment difficult. Wigwams was used to introduce the network footprint components into the corresponding transcription factor-only models, resulting in enhanced network models carrying information about downstream regulated genes. This allows for functional assessment to be used to identify nodes of interest within the network, and propose concise follow-up experiments.

Abbreviations

ABA	abscisic acid
ABRE	ABA-responsive element
BATS	Bayesian Analysis of Time Series
bHLH	basic helix loop helix
bp	base pair
bZIP	basic leucine zipper
cfu	colony forming unit
ChIP	chromatin immunoprecipitation
CSI	Causal Structure Inference
CUC	cup-shaped cotyledon
das	days after sowing
DBD	DNA binding domain
DC3000	*Pseudomonas syringae* pv. *tomato* DC3000
DE	differentially expressed
DEG	differentially expressed gene
DNA	deoxyribonucleic acid

EAR	ethylene responsive element binding factor associated amphiphilic repression
EBS	EIN3 binding site
ECTDISA	Enrichment Constrained Time-Dependent Iterative Signature Algorithm
EDISA	Extended Dimension Iterative Signature Algorithm
EM	Expectation Maximisation
ER	endoplasmatic reticulum
ERF	ethylene response factor
ETI	effector triggered immunity
ETS	effector triggered susceptibility
FDR	false discovery rate
GGM	graphical Gaussian model
GO	Gene Ontology
GP	Gaussian process
GP2S	Gaussian process two-sample test
GTF	general transcription factor
GUI	graphical user interface
HKDM	histone lysine demethylase
HKMT	histone lysine methyltransferase
hpi	hours post inoculation
HR	hypersensitive response
HRP	hypersensitive reaction and pathogenicity

ISA	Iterative Signature Algorithm
JA	jasmonic acid
JA-Ile	jasmonoyl-L-isoleucine
JAZ	jasmonate ZIM-domain-containing
LIMMA	Linear Models for Microarray Data
LRR	leucine-rich repeat
LRR-RK	leucine-rich repeat receptor kinase
LTAE	long-term agroecosystem experiment
M-VBSSM	Metropolis VBSSM
MAANOVA	Microarray Analysis of Variance
MAMP	microbe-associated molecular pattern
MAPK	mitogen-activated protein kinase
MEME	Multiple EM for Motif Elongation
mRNA	messenger RNA
MSR	mean square residue
MTI	MAMP triggered immunity
NAC	NAM, ATAF, CUC
NAM	no apical meristem
NB	nucleotide-binding
ODE	ordinary differential equation
PAMP	pathogen-associated molecular pattern

PBM	protein-binding microarray
PCC	Pearson correlation coefficient
PDI	protein disulfide isomerase
PP2C	protein serine/threonine phosphatase 2C
PR	pathogenesis-related
PRR	pattern recognition receptor
PSSM	position specific scoring matrix
PTI	PAMP triggered immunity
pv.	pathovar
R gene	resistance gene
RING	Really Interesting New Gene
RLK	receptor like kinase
RLP	receptor like protein
RNA	ribonucleic acid
ROC	receiver operating characteristic
ROS	reactive oxygen species
SA	salicylic acid
SNRK2	SNF-1 related protein kinase 2
SSM	state space modelling
T-DNA	transfer DNA
T3SS	type III secretion system

TAF	TBP associated factor
TAIR	The Arabidopsis Information Resource
TBP	TATA binding protein
TF	transcription factor
TFIIA	transcription factor II A
TFIIB	transcription factor II B
TFIID	transcription factor II D
TFIIE	transcription factor II E
TFIIH	transcription factor II H
TPL	TOPLESS
TR	transcriptional repressor
TRF	TBP related factor
TTSS	type III secretion system
UV	ultraviolet
VBSSM	Variational Bayesian State Space Modelling
Wigwams	Wigwams identifies genes working across multiple situations
Y1H	yeast one-hybrid
Y2H	yeast two-hybrid

Gene Names

ABI5	AT2G36270
AFP3	AT3G29575
ANAC019	AT1G52890
ANAC055	AT3G15500
ANAC058	AT3G18400
ANAC072	AT4G27410
ANAC100	AT3G15170
ARF3	AT2G33860
ATAF1	AT1G01720
BAK1	AT4G33430
BOI	AT4G19700
BOS1	AT3G06490
bZIP63	AT5G28770
CCA1	AT2G46830
COI1	AT2G39940
CRK45	AT4G11890

CTR1	AT5G03730
CUC1	AT5G61430
EBF1	AT2G25490
EBF2	AT5G25350
EDS1	AT3G48090
EFR	AT5G20480
EIL1	AT2G27050
EIN2	AT5G03280
EIN3	AT3G20770
ERF1	AT3G23240
ERF7	AT3G20310
ERF8	AT1G53170
FBH3	AT1G51140
FLS2	AT5G46330
FRK1	AT2G19190
GRP7	AT2G21660
GRX480	AT1G28480
HDA6	AT5G63110
HDA19	AT4G38130
JAR1	AT2G46370
LHCA6	AT1G19150

LHY	AT1G01060
LOX2	AT3G45140
MAP kinase 4	AT4G01370
MAPKKK18	AT1G05100
MIN7	AT3G43300
MYC2	AT1G32640
NINJA	AT4G28910
NPR1	AT1G64280
ORA59	AT1G06160
PAD4	AT3G52430
PBS1	AT5G13160
PDF1.2	AT5G44420
PDI1	AT3G54960
PDI2	AT5G60640
PDI5	AT1G21750
PDI6	AT1G77510
PEN2	AT2G44490
PEN3	AT1G59870
PIL5	AT2G20180
PKS5	AT2G30360
PNSL2	AT1G14150

PR1	AT2G14610
PSB27	AT1G03600
PSB29	AT2G20890
PSBO1	AT5G66570
PSBQA	AT4G21280
PUB22	AT3G52450
PUB23	AT2G35930
Rap2.6	AT1G43160
RAV1	AT1G13260
RAV2	AT1G68840
RBOHD	AT5G47910
RIN4	AT3G25070
RPM1	AT3G07040
RPS2	AT4G26090
RPS4	AT5G45250
RPS5	AT1G12220
RRS1-R	AT5G45260
SAP12	AT3G28210
SAP18	AT2G45640
SDG26	AT1G76710
SDG33	AT5G13960

SIGA	AT1G64860
STZ	AT1G27730
SWI3B	AT2G33610
TBP3	AT5G67580
TCP1	AT1G67260
TCP3	AT1G53230
TCP20	AT3G27010
TGA3	AT1G22070
TOC1	AT5G61380
TOPLESS	AT1G15750
VSP2	AT5G24770
WIN2	AT4G31750
WRKY9	AT1G68150
WRKY11	AT4G31550
WRKY17	AT2G24570
WRKY22	AT4G01250
WRKY29	AT4G23550
WRKY33	AT2G38470
WRKY38	AT5G22570
WRKY41	AT4G11070
WRKY45	AT3G01970

WRKY53	AT4G23810
WRKY62	AT5G01900
WRKY70	AT3G56400
XND1	AT5G64530
ZAR1	AT3G50950
ZED1	AT3G57750

Chapter 1

Introduction

1.1 Motivation

1.1.1 Feeding an exponentially growing population

Food security remains one of the primary challenges in need of tackling by current society, predominantly as a result of the constant, exponential increase in the world population. According to the United States Census Bureau, the world population crossed 7 billion people on March 12, 2012. This figure has been conservatively estimated to reach 9.5 billion by 2050 (United Nations, 2012), marking an increase of 2.5 billion people over the span of 38 years. This creates the need to optimise food production yield, as 15 million km^2 are already devoted to the growth of crops worldwide (Ramankutty et al., 2008) and any additional expansion of this land is likely to involve taking over space currently occupied by rain forests and other areas with significant ecological functionality (Gibbs et al., 2010).

Food security has improved dramatically over the course of the past century, predominantly due to the introduction of chemical compounds capable of combatting the common pests adversely affecting the crops. The longest running long-term agroecosystem experiment (LTAE) at Rothamsted had a tremendous increase in crop yield with the introduction of pesticides, and subsequently fungicides (Rasmussen et al., 1998). However, in a similar scenario to bacteria developing resistance to antibiotics, the effectiveness of pesticide use is slowly wearing out. By now, resistance to one or more pesticides has been documented in over 400 insects or mites (Roush and Tabashnik, 2012). That, in combination with the numerous drawbacks of pesticides, including the contamination of soil, ground water and surface water, negative impact on soil fertility and adverse effects on non-target organisms including humans (Aktar et al., 2009), implies that a preferred long-term strategy would

not involve this sort of treatment. In addition to the complications from the most popular methods of handling biotic stimuli, a number of abiotic environmental conditions are also problematic to crop yield. The most prevalent of these is drought — worldwide, it is responsible for crop losses up to 60% (Bruce et al., 2002). Only 16% of all farmland gets the appropriate amount of water (Siebert et al., 2005) and climate change is likely to exacerbate the drought problem (Wheeler and von Braun, 2013).

A natural course of action is to investigate the nature of the plant's response to such stimuli, leading to the ability to augment the stimulus response and increase resilience. It should be noted that attempts at plant modification to decrease susceptibility to biotic or abiotic stimuli might adversely affect crop yield. As evidenced by the maize mutant *opaque-2*, increasing resilience towards stress can come at the cost of loss of biomass, as the plant is tuned towards damage prevention and survival (Herms and Mattson, 1992). Nevertheless, improvements in the existing farming infrastructure to decrease crop exposure to stress combined with careful enhancement of the plant's natural ability to handle adverse conditions should result in increased resilience and steady yield, whilst lessening the reliance on chemical agents.

1.1.2 *Arabidopsis thaliana* as a model organism in studying plant stress response

Thale cress (*Arabidopsis thaliana*) has been the model organism of choice in plant research. *A. thaliana* possesses one of the smallest plant genomes, spanning 115.4 Mbp across five chromosomes. It was the fourth published genome sequence (Arabidopsis Genome Initiative, 2000), preceded only by baker's yeast (*Saccharomyces cerevisiae*, the model eukaryote) (Goffeau et al., 1996), the nematode (*Caenohabditis elegans*) (C. elegans Sequencing Consortium, 1998) and the fruit fly (*Drosophila melanogaster*) (Adams et al., 2000). The subsequent creation of The Arabidopsis Information Resource (TAIR) (Rhee et al., 2003) allowed for the easy integration of a variety of information sources, including extensive gene annotations and the positioning of T-DNA insertions in mutant lines, centralising the access to a high amount of information critical for research.

A. thaliana is also easy to mutate, allowing for the assessment of the role of a select group of genes in a particular process by subjecting a knockout or overexpressor to the experimental conditions of interest. The mutation procedure involves random T-DNA insertion performed by *Agrobacterium tumefaciens*, with the initial developed procedure introducing *A. tumefaciens* to the plant through an incision at the base of the apical shoots (Chang et al., 1994). The method was improved with

the development of floral dip, wherein developing *A. thaliana* floral tissue is dipped in an *A. tumefaciens*, sucrose and surfactant solution (Clough and Bent, 1998). A large-scale knockout experiment was performed by SALK, in which a conservative set of 88,122 long T-DNA integration sites were identified to span almost 75% of the *A. thaliana* genes (Alonso et al., 2003). The great degree of success of this approach stems from the high gene density of *A. thaliana*, with its genome containing low amounts of intragenic DNA and being of overall small size for its gene count (Feuillet and Keller, 1999). Other T-DNA insertion collections, such as SAIL (Sessions et al., 2002) and GABI-KAT (Kleinboelting et al., 2011) have also been developed. The resulting lines are stored in stock centres, such as ABRC and NASC, all over the world for ease of access. It is also possible to obtain constitutive overexpressors by introducing a fusion of the gene of interest under the control of the cauliflower mosaic virus 35S promoter to the genome (Onouchi et al., 2000). A more refined control over the expression of a gene of interest can be achieved through the use of promoters stimulated through an exposure to a particular compound, such as dexamethasone (Ullah et al., 2001).

Other advantages of *A. thaliana* involve its practicality from an experiment design perspective. The life cycle of *A. thaliana* is very rapid, going from seed to seed in 5 to 7 weeks, allowing for the quick establishment of mutations and elucidation of lines with multiple genes knocked out (Boyes et al., 2001). It is possible to grow *A. thaliana* in many different environments, such as greenhouses, fluorescent lights in laboratory incubators or petri dishes, allowing for flexible and versatile experimental application (Meinke et al., 1998). The plants are small, with full-grown Col-0 rosettes reaching an average size of 8 cm by 7 cm. Plants can be conveniently grown next to each other. In addition to that, *A. thaliana* has a high seed yield, averaging about 5000 seeds per plant (Boyes et al., 2001). This allows for a single genotype to be easily maintained throughout the course of an experiment. Additionally, the seed stock can be easily and compactly stored in a cold room.

Due to the diversity in the plant kingdom, research findings made on *A. thaliana* may not be immediately transferable to species with more real-world application, but the general sense of the findings should be preserved. In spite of the fact that *A. thaliana* is a dicot and a number of the more popular crop plants, such as rice, wheat and maize, are monocots, a degree of conservation of genetics and physiology is present between all plant species. This is manifested by numerous examples of expression alteration of a gene giving similar phenotypic results between *A. thaliana* and rice (Flavell, 2009). In addition, isologs and paralogs of numerous crop plant stress responsive genes have been identified in *A. thaliana* (Zhu,

2000), further backing the claim of a conserved inter-species nature and asserting the validity of *A. thaliana* as a model organism for studying stress response.

1.2 Stress response in *A. thaliana*

Plants regularly encounter a variety of both biotic and abiotic environmental conditions adversely affecting their development and survival, henceforth referred to as stresses. Abiotic stresses, such as drought, high light, cold, high salinity and UV light exposure, are far more prevalent than pathogen infections and account for a steady loss of yield in crops worldwide (Kreps et al., 2002). Whilst less prevalent than abiotic stresses, pathogen infections are also of agricultural significance. This form of stress will be the main focus of the following sections due to the scope of the work carried out in Chapters 2, 3 and 4.

In spite of the lack of adaptive immunity and a vast array of pathogenic organisms present in the environment, plants are able to resist the attacks of the majority and are usually only susceptible to a select few. This stems from the depth of the plant defence response, which starts at preformed physical and chemical layers designed to keep the pathogens out, and is optionally followed by a breadth of inducible responses (including defence-related gene expression, production of antimicrobial compounds, oxidative bursts and programmed cell death) in case a pathogen manages to breach the preformed layers (Van Loon et al., 2006). The plant defence system, which is outlined in Figure 1.1 (Jones and Dangl, 2006), starts at host encoded pattern recognition receptors (PRRs), which are capable of identifying a range of pathogen-associated molecular patterns (PAMPs), which are conserved elements of compounds that the pathogens need for basic functionality, such as flagellin (Zipfel and Felix, 2005). This triggers a response known as PAMP triggered immunity (PTI), which is sufficient to prevent infection by many pathogens and likely forms the core of non-host resistance. A number of successful pathogens are capable of secreting specific proteins, called effectors, into the plant, with the aim of suppressing PTI and facilitating disease through active manipulation of the plant's cellular processes. This is known as effector triggered susceptibility (ETS). In response to that, the plant is capable of direct or indirect identification of effectors and activating another layer of defence to counteract them, resulting in disease resistance. This is known as effector triggered immunity (ETI) (Jones and Dangl, 2006).

Activation of the plant stress response is reliant on a number of hormones, with four of the primary ones being jasmonic acid (JA), salicylic acid (SA), abscisic

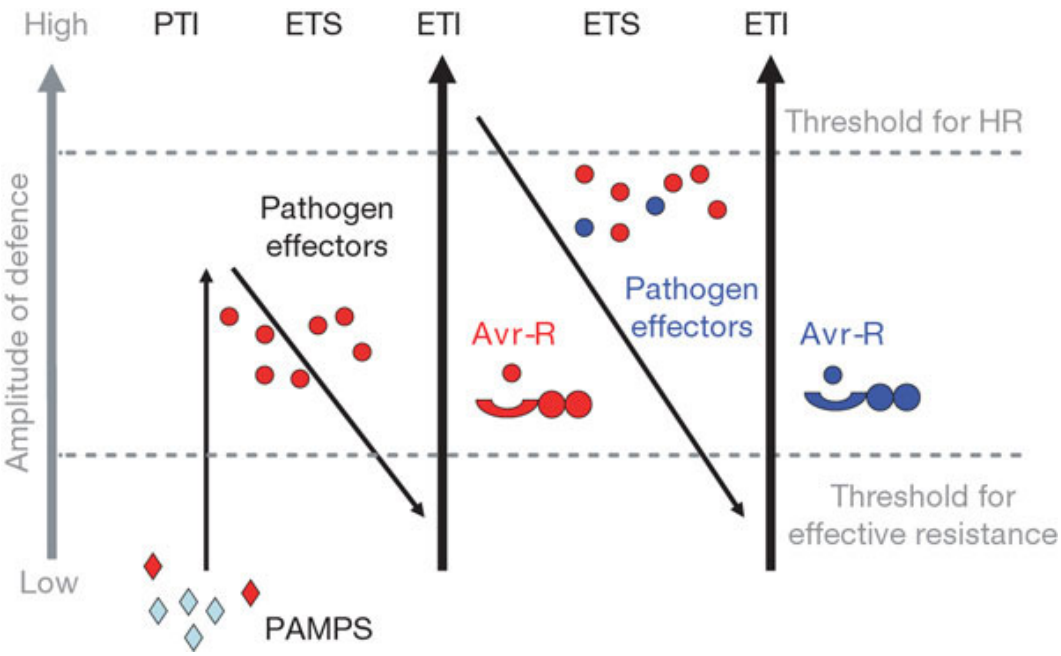

Figure 1.1: The zigzag model, capturing the interactions between the plant and the pathogen. The initial defence response, known as PTI, is triggered through the detection of conserved compounds the pathogen relies on for basic functionality. The pathogen can suppress this response by introducing effectors into the plant, resulting in ETS. The plant can in turn detect effector activity within its cells, and initiate ETI in response. This is a constant evolutionary arms race between the two organisms, with pathogens selected for those with novel effectors leading to disease, and in turn plants being selected for those capable of detecting the altered defence suppression. Reproduced from Jones and Dangl (2006).

acid (ABA) and ethylene. Signalling induced by these hormones, as well as the resulting crosstalk between them, is a key element of the response to a variety of biotic and abiotic stresses (Bari and Jones, 2009). Effects of this signalling, as well as from other signalling pathways triggered by the stress, result in a massive transcriptional overhaul — both biotic (Windram et al., 2012) and abiotic (Kilian et al., 2007) stimuli result in the altering of expression of thousands of genes with the aim of responding to the stimulus. In the case of abiotic stress, there appears to be an early conserved condition-independent general transcriptional response that subsequently diverges into more stress-specific processes. This was first identified for cold, salt and osmotic stress, with a core of 65 genes showcasing joint differential expression across all three conditions at 3 hours of exposure, with the overlap falling significantly by 27 hours of exposure (Kreps et al., 2002). A similar early non-specific response was later observed across cold, drought and UV-B light, hinting at a core of plant environmental stress response genes (Kilian et al., 2007).

A common environmental scenario is the co-occurrence of multiple biotic and abiotic stimuli, often demanding antagonistic action from the plant that it is unable to perform (Mittler, 2006). The most prevalent example of this would be the joint application of drought and heat stress. In the case of heat, the plant's primary response is to open its stomata to cool the leaves by transpiration. However, in the case of drought, opening stomata is repressed. As such, a combination of the two is disastrous, where the plant has to cope with the effects of drought whilst at the same time having its leaf temperature increased by the heat (Rizhsky et al., 2002). This results in massive crop losses — in August 2000 alone, a co-occurrence of drought and heat stress across the USA caused losses upwards of $4.2 billion (Mittler, 2006). This signals an important avenue for follow-up research once a sufficient depth of understanding of individual stress responses is reached, due to the lack of realism of a single isolated stress occurring naturally.

1.2.1 PAMP triggered immunity

The first wave of inducible pathogen defence response is mediated by pattern recognition receptors (PRRs), which are transmembrane proteins capable of the identification of slowly evolving pathogen-associated molecular patterns, such as flagellin. Flagellin is a key component of the bacterial flagellum, the main role of which is locomotion (Zipfel and Felix, 2005). A 22 amino acid fragment of a conserved flagellin domain, synthesised under the name flg22, is sufficient to trigger a widespread, rapid response in *A. thaliana* (Zipfel et al., 2004). It has been shown that the receptor binding flagellin in *A. thaliana* is a leucine-rich repeat receptor kinase

(LRR-RK) FLS2 (Gømez-Gømez and Boller, 2000). FLS2 binds flg22 with its LRR ectodomain, changing spatial conformation. This allows BAK1, another LRR-RK that is co-located with FLS2 but unable to interact with it in the absence of a flagellin stimulus, to create a complex with FLS2. The newly formed heterodimer brings the intracellular kinase domains into contact, initiating defence signalling (Boller and Felix, 2009). It has been shown that knocking out FLS2 in *A. thaliana* leads to an increased susceptibility to the pathogen *Pseudomonas syringae* pv. *tomato* DC3000 when it is sprayed onto the leaf surface (Zipfel et al., 2004) presumably through a reduced ability to activate PTI. Plants are not restricted to identifying a single PAMP per pathogen, with the identification of bacteria not merely relying on flagellin, but also aided by the detection of other compounds, such as EF-Tu (Zipfel et al., 2006) A number of PAMPs, representative of a number of different pathogens, are bound by known PRRs, including a glucan receptor in soybean (Fliegmann et al., 2004), a chitin receptor in rice (Kaku et al., 2006) and *A. thaliana* (Miya et al., 2007), and a xylanase receptor in tomato (Ron and Avni, 2004). Additional PAMPs are well defined, but the current state of PRR knowledge is insufficient to assign particular LRR-RKs to them (Boller and Felix, 2009).

The predominant role of PTI is to halt further pathogen colonisation (Jones and Dangl, 2006). The earliest effects, observed within 5 minutes of the stimulus, include ion fluxes leading to membrane depolarisation (Mithöfer et al., 2005), and an activation of the mitogen-activated protein kinase (MAPK) cascade culminating in the activation of WRKY-type transcription factors (Asai et al., 2002), reactive oxygen species (ROS) production (Asai et al., 2008), and an increase in protein phosphorylation (Boller and Felix, 2009). The next wave of effects, observed up to 30 minutes after the stimulus, are the biosynthesis of ethylene (Spanu et al., 1994), retraction of PRRs (possibly related to the degradation of activated receptors) (Robatzek et al., 2006), and major transcriptional changes (Zipfel et al., 2004). Long-term effects are callose deposition (Gømez-Gømez et al., 1999), and the inhibition of growth as the plant shifts its physiology to a defence role (Boller and Felix, 2009).

1.2.2 Effector triggered susceptibility

From the perspective of a pathogen attempting to infect the plant, the PTI response is undesirable and would be preferably avoided. PRRs target molecular signatures that are essential to pathogen survival and hence not easy to shed. To circumvent this, some microbes have evolved a strategy involving the introduction of effector proteins, which are capable of altering the signalling and metabolism within the

plant to derail PTI or make the infection process easier (Jones and Dangl, 2006). Different pathogen species secrete effectors in alternate ways — in bacteria, effectors are introduced to the plant through the means of the type III secretion system (TTSS), taking on the form of a needle protruding from the surface of the pathogen (Kubori et al., 1998), whilst oomycetes and fungi utilise haustoria, finger-like structures that protrude into the plant cell and secrete effectors that subsequently cross through the cell membrane with the aid of translocation domains (Bozkurt et al., 2012).

Once present in the plant cell, the nature of the tasks carried out by the effectors is quite varied, can involve both up- and down-regulation, and only a fraction of their targets have been identified. Two *P. syringae* DC3000 effectors, AvrPto and AvrPtoB, are capable of forming complexes with the kinase domain of BAK1 (Shan et al., 2008), making it impossible for the BAK1-FLS2 complex outlined in the previous section to be created in response to flagellin detection, effectively muting the PRR and derailing the PTI response. Additionally, AvrPtoB was found to contain a domain similar in structure to E3 ubiquitin ligase, resulting in the plant cell degrading the PRR components the effector binds to (Rosebrock et al., 2007). HopAI1, another *P. syringae* effector, interrupts the PRR signalling at a later stage by dephosphorylating kinases of the MAPK cascade (Zhang et al., 2007). Effectors can also target processes downstream of the initial PRR response, halting phenomena that make the infection process more difficult. HopU1 modifies a number of *A. thaliana* RNA-binding proteins, with GRP7 as an example, whilst HopM1 triggers the degradation of MIN7, a protein involved in vesicle trafficking. Knockout lines for the two aforementioned genes showed increased susceptibility to *P. syringae* infection, indicating a role of vesicle trafficking and RNA-binding in the defence response (Block et al., 2008). Another *P. syringae* effector, HopI1, prevents the accumulation of salicylic acid, a hormone with a key role in the biotrophic pathogen defence response (Boller and He, 2009). A key effector target is RIN4, which has been shown to have an important role in regulating stomatal opening (Liu et al., 2009), a port of easy entrance for the pathogen into the leaf. Three distinct effectors target this protein — AvrRpt2 is a cysteine protease that activates once inside the plant cell (Coaker et al., 2005) and cleaves RIN4 at two sites (Chisholm et al., 2005), whilst AvrRpm1 and AvrB phosphorylate RIN4, rendering it inactive (Mackey et al., 2002).

1.2.3 Effector triggered immunity

Shifting back to the plant perspective, the pathogen injecting effectors and phytotoxins that result in the hijacking of the defence response is detrimental. As such, plants evolved the ability to respond to the effector actions, triggering a second wave of defence response deemed effector triggered immunity (ETI) (Jones and Dangl, 2006). R (resistance) genes are protein receptors located within the plant cell responsible for detecting effectors and/or their activity. There is a degree of variation in the structure and domains of R genes, but the two components that are always present are a nucleotide-binding (NB) domain and leucine-rich repeat (LRR) domain. The mechanics of the identification of effector action involve a bait and switch model, wherein the R gene protein is inactive until the detection of the signal by the joint work of the LRR domain and variable N-terminus domains results in a spatial conformation shift that activates the NB domain, initiating the defence response (Collier and Moffett, 2009). In the case of the pathogen targeting RIN4, as outlined in the previous section, *A. thaliana* possesses two distinct R genes that detect the presence of the effectors indirectly. RPS2 identifies the pieces of RIN4 that were cleaved by AvrRpt2 (Mackey et al., 2003), whilst RPM1 detects RIN4 phosphorylated by AvrRpm1 or AvrB (Mackey et al., 2002). R genes are not limited to indirect detection of effector activity — for example, RRS1-R directly binds the effector PopP2, leading to *Ralstonia solanacearum* resistance (Deslandes et al., 2003).

It should be of little surprise that the actual mechanics of ETI response are relatively similar to PTI — PTI is quite potent at handling pathogens, so ETI activates those processes again, getting around the pathogen's attempts at not being detected. ROS production and MAPK cascade activation return, but at a much higher intensity and prolonged duration (Tsuda and Katagiri, 2010). A massive transcriptional reprogramming is also observed, with WRKY transcription factors once again being a key driving force behind the overhaul (Wu et al., 2014). A process not as common in PTI is the hypersensitive response (HR), leading to rapid programmed cell death restricting pathogen growth. Some of its triggers are the detection of AvrRps4 by R gene RPS4, or the detection of AvrRpm1 by RPM1 (Aarts et al., 1998). It has been shown that autophagy, a catabolic mechanism leading to the degradation of unnecessary cell components, plays an important role in carrying out the programmed cell death (Hofius et al., 2009).

An interesting plant-side evolutionary strategy involves the modification of effector targets into non-functional decoys, allowing for the easy identification of pathogen infection without impairing the defence response. An example of this is

ZED1, a non-functional kinase, which is acetylated by HopZ1a and subsequently triggers ETI via ZAR1 (Lewis et al., 2013). The arms race between the plant and the pathogen, wherein the plant attempts to get rid of the pathogen and the pathogen attempts to stifle the defence response and carry out infection, has been named the zigzag model. This is an on-going process, as pathogen selection can lead to changes in the secreted effectors, avoiding triggering ETI. This results in plant selection, with individuals capable of identifying the new effectors via novel R genes being favoured (Jones and Dangl, 2006).

1.2.4 Examples of plant pathogens

During its life, a plant is likely to encounter a number of biotic stimuli. Outlined below are two pathogens, *Pseudomonas syringae* and *Botrytis cinerea*, representative of hemibiotrophic bacteria and necrotrophic fungi respectively.

1.2.4.1 *Pseudomonas syringae*

Pseudomonas syringae is a Gram-negative rod bacterium with polar flagellation. The *syringae* stems from *Syringa vulgaris* (lilac), the plant species from which the original *P. syringae* strain was isolated from by van Hall in 1902 (Hirano and Upper, 1990). *P. syringae* is hemibiotrophic, initially living on the exterior of the plant as an epiphyte, and later moving into the apoplast (intracellular space) as a pathogenic endophyte (Jin et al., 2003). The ability to cause disease by a particular *P. syringae* strain is dependent on the host, and as such *P. syringae* strains are grouped into pathovars (pv.) by host range (Block and Alfano, 2011). In the case of disease, the prime outcome is plant tissue necrosis, which leads to symptoms ranging from leaf spots to stem cankers (Jin et al., 2003). The early availability of the genomes of a number of *P. syringae* strains, in combination with their ability to cause disease in model plant organisms, has led *P. syringae* to become the model system for studying plant-pathogen interactions (Mansfield, 2009).

A key component of the *P. syringae* infection process that happens in the apoplast is the injection of effectors through the type III secretion system (TTSS) in order to combat PTI, following the typical pathogen-plant interaction as previously discussed with numerous *P. syringae* examples in sections 1.2.1 through 1.2.3 (Jones and Dangl, 2006). The *P. syringae* type III secretion system is constructed of hypersensitive reaction and pathogenicity (*HRP*) genes (Lindgren et al., 1986), with their expression only induced when the bacteria were placed in plant tissue or in a medium mimicking the conditions present in the apoplast (Rahme et al., 1992).

The *HRP* genes can be divided into three classes — *HRC* (*HRP* genes conserved), core TTSS structural components that share a great deal of sequence similarity with flagellum assembly genes, implying a common ancestry of both structures (Gophna et al., 2003); regulatory proteins that induce the expression of the remaining *HRP* genes in apoplast-like environmental conditions (Bretz et al., 2002); and some of the secreted effectors and extracellular structural components of the TTSS (Jin et al., 2003). This includes HrpA, one of the aforementioned extracellular structural components, with a pivotal role in the creation of the effector-secreting needle — the deletion of this gene results in an avirulent *P. syringae* strain, incapable of delivering effectors to the plant (Roine et al., 1997). In the case of successful introduction of effectors into the plant in an attempt to derail the defence response and adjust any desired cellular processes, it is possible for the plant to detect effector presence through R genes and initiate ETI. Examples of *A. thaliana P. syringae*-specific R genes include RPM1, RPS2, RPS4, RPS5 and PBS1 (Katagiri et al., 2002). It should be noted that *P. syringae* does not just utilise effectors during infection — a number of phytotoxins are also secreted, including coronatine, which mimics jasmonic acid and overcomes salicylic acid-based biotic pathogen defence responses (Brooks et al., 2005), with its main role being overcoming stomatal closure initiated as part of the defence response (Melotto et al., 2006), and syringolin, which inhibits the proteasome (Groll et al., 2008).

P. syringae is also capable of ice nucleation, with the bacteria inducing frost damage on plants in temperatures higher than unassisted environmental conditions, potentially in order to access nutrients (Hirano and Upper, 1990). Subsequently, the entire life cycle of the bacteria was identified, with the ice nucleation ability playing a relevant role in precipitation — epiphytic *P. syringae* on plant surfaces get taken up to the clouds by air currents, from which they are deposited to lakes and rivers by rain and snow, from where they are transported to plants via either natural or human-created irrigation (Morris et al., 2008).

1.2.4.2 *Botrytis cinerea*

Botrytis cinerea is an airborne necrotrophic fungal pathogen, with its strains capable of causing grey mould to develop on leaves, shoots, flowers, fruits and storage organs of over 200 primarily dicot crop species, including carrots, cabbage, broccoli, grapes and strawberries (Droby and Lichter, 2007). *B. cinerea* conidia (spores) typically gain entry to the crop whilst it is still on the field, but only start germinating and infecting the plant once the tissues are senescent. This can result in heavy post-harvest losses when an apparently healthy crop gets harvested and only begins

exhibiting symptoms in transport or when placed on sale (Williamson et al., 2007). The established way to combat *B. cinerea* is with the aid of a variety of fungicides, but the pathogen has begun exhibiting resistance to some of the compounds in use (de Waard et al., 2006).

Once *B. cinerea* senses that the host tissue is sufficiently rich in nutrients, the conidia germinate and form appressoria, a specialised infection structure that breaches the host cuticle with the aid of a penetration peg (van Kan, 2006). The penetration pegs don't rely on pressure alone to make their way into the plant, but are aided by cutinases and lipases when breaching the cuticle and by pectinases and cellulases when reaching the plant cell wall underneath (Williamson et al., 2007). Once the plant senses the infection through chitin detection, it initiates the PTI response (Mengiste, 2012). As part of the PTI response, the plant secretes ROS in an attempt to get rid of the pathogen, which the fungus responds to by secreting catalases and peroxidases that decompose the ROS (Schouten et al., 2002). *B. cinerea* proceeds to kill the plant cells, utilising phytotoxic proteins inducing cell death (Staats et al., 2007), botrydial and other phytotoxic metabolites (Colmenares et al., 2002) and oxalate, which has been shown to cause plant wilting (Dutton and Evans, 1996).

The *A. thaliana* response to *B. cinerea* infection is considerably different to that following *P. syringae* infection, as the main goal of the fungal pathogen is inducing necrosis and the plant has to attempt to counteract the phytotoxic effects of the introduced compounds. The hypersensitive response, which is a component in the response to *P. syringae* infection, is actually of huge benefit to *B. cinerea* (Govrin and Levine, 2000). Another divergence from the response to the aforementioned bacterial pathogen is the lack of R gene action — instead of a full ETI activation if any of the gene-for-gene relationships are satisfied, *B. cinerea* resistance is conferred incrementally through the action of multiple genes. Examples of genes involved in this incremental resistance include the transcription factor BOS1, potentially activated through ROS emitted by the pathogen, with its downstream signalling aiming to inhibit necrosis (Mengiste et al., 2003), and the RING E3 ubiquitin ligase BOI restricting toxin-induced cell death triggered by α-picolinic acid (Mengiste, 2012).

It should also be noted that in spite of most *B. cinerea* infections being adverse to agriculture, it also possesses a favourable use. An application of *B. cinerea* to wine grapes under moist, favourable conditions, but with the weather becoming dryer soon thereafter, results in a specific form of infection known as 'noble rot', with the affected grapes being used in wine making (Rosslenbroich and

Stuebler, 2000).

1.2.5 Phytohormones in the plant stress response

1.2.5.1 Jasmonic acid

Jasmonic acid (JA) is a lipid-derived molecule, originating from the fatty acid α-linoleic acid from chloroplast membranes (Delker et al., 2006). The role of JA is very broad, with the phytohormone being involved in the regulation or signalling of developmental functionality, senescence, herbivore attack, wounding, light response and abiotic stress, but its primary role is in the response to necrotrophic pathogens (Bari and Jones, 2009; Robert-Seilaniantz et al., 2011). JA-induced signalling leads to the production of assorted stress-related metabolites, including glucosinolates, phenylpropanoids and anthocyanins (Pauwels and Goossens, 2008).

The JA signalling pathway involves the hormone molecule initiating the degradation of jasmonate ZIM-domain-containing (JAZ) transcriptional repressors. In the absence of JA, JAZ proteins bind target transcription factors, forming a complex with TOPLESS (TPL) mediated by NINJA (Pauwels and Goossens, 2008) that represses the activity of the bound transcription factor through histone deacetylation (Causier et al., 2012). The 12 JAZ proteins are capable of homo- and heterodimerising *in vitro* through a conserved TIFY domain, suggesting a role in the fine-tuning of the signal response (Chini et al., 2009). In the presence of JA, JAR1 conjugates it with isoleucine forming JA-Ile (Staswick and Tiryaki, 2004), leading JAZ proteins to form a complex with COI1, which subsequently gets degraded (Sheard et al., 2010). This results in the decomposition of the structure repressing the transcription factor targets, allowing them to induce transcription of relevant genes. The transcription factors bound by JAZ proteins can induce two distinct response pathways, with MYC2 predominantly leading to the wounding response (Pieterse et al., 2009), whilst EIN3 and EIL1 drive the defence response, with key downstream targets including the transcription factors ERF1, PDF1.2 and ORA59 (Zhu et al., 2011).

1.2.5.2 Salicylic acid

Salicylic acid (SA) is a monohydroxybenzoic acid with a similarly broad role in plant signalling to JA. SA has been shown to be involved in flowering, fruit ripening, seed germination, seedling establishment and response to abiotic stresses. The hormone's primary function is in response to biotrophic pathogens, in particular in the induction of the hypersensitive response (HR), leading to rapid programmed

cell death, and the development of systemic acquired resistance (SAR) (Vlot et al., 2009). SAR is a phenomenon wherein plant tissue located away from the primary pathogen infection site develops long-term infection resistance, primarily through the induction of pathogenesis-related (PR) genes (Durrant and Dong, 2004). An example PR gene is PR1, which confers resistance to *P. syringae* by limiting pathogen growth (Glazebrook, 2005). SA is also an active metabolite of aspirin.

SA synthesis is triggered by EDS1 and PAD4 upon sensing pathogen presence (Durrant and Dong, 2004). The produced SA spreads to distal plant cells, wherein it activates NPR1. In the absence of SA, NPR1 forms oligomers in the cytoplasm, but upon SA introduction, the resulting redox change splits the NPR1 complex into monomers and the proteins get transported into the nucleus (Mou et al., 2003). Once there, NPR1 binds transcription factors from the TGA and WRKY families and leads to the induction of PR gene transcription (Pieterse et al., 2009).

1.2.5.3 Abscisic acid

Abscisic acid (ABA) is an isoprenoid, obtained in plants by cleaving 40-carbon carotenoids (Nambara and Marion-Poll, 2005). The hormone's name derives from the past belief that it plays a key role in the abscission (shedding) of leaves, but it has since been shown to not be the case (Jackson and Osborne, 1970). ABA is only involved in regulating abscission in a small number of plant species (Gomez-Cadenas et al., 2002). The hormone plays a role in a variety of developmental processes, such as seed germination, dormancy, lateral root formation, flowering inhibition and senescence, as well as a number of abiotic stresses (Asselbergh et al., 2008). ABA's role in the abiotic stress response is predominantly centred on plant water balance, including inducing stomatal closure in drought (Tuteja, 2007). Assessing the transcriptional response of rice to cold, drought, salt and ABA treatment revealed a strong association between the induced genes, in particular between drought, salt and ABA exposure, further cementing ABA's role as a crucial agent in the abiotic stress response (Rabbani et al., 2003). Interestingly, ABA's involvement in response to pathogen infection is a forceful regulation of a negative defence phenotype, going as far as inducing disease symptoms from *Cladosporium cucumerinum*, a potato nonpathogen, on ABA-treated potato slices (Henfling et al., 1980). The role of ABA in pathogen defence response appears to be pathogen-specific, though, as *A. thaliana* strains with mutations in the ABA biosynthesis genes were found to be more susceptible to a number of necrotrophic pathogens, including *Pythium irregulare* (Adie et al., 2007).

ABA signalling is similar in nature to JA, as it induces the activity of an

otherwise repressed protein. In the absence of ABA, SNF-1 related protein kinase 2s (SnRK2s) are bound and inactivated by protein serine/threonine phosphatase 2Cs (PP2Cs). Upon ABA introduction, the hormone is bound by cytosolic receptors from the PYL family. Forming a complex with ABA alters the conformation of PYLs, allowing them to bind PP2Cs, inhibiting their active site, leading to the activation of SnRK2s through alleviation of negative regulation (Cutler et al., 2010). Consistently with the previously discussed hormones, ABA response involves massive transcriptional reprogramming, mediated by transcription factors that bind ABRE (ABA-responsive elements) in the promoters of target genes (Uno et al., 2000). This is predominantly carried out by AREB/ABF transcription factors, members of the bZIP transcription factor superfamily, which are activated through phosphorylation by SnRK2s (Fujita et al., 2013).

1.2.5.4 Ethylene

Ethylene is the simplest alkene, taking on the form of gas. In plants, it controls a variety of growth and development processes. The most notable of these are senescence-related, with fruit ripening being the most agriculturally relevant (Bleecker and Kende, 2000). Ethylene was also found to be the primary leaf abscission hormone (Jackson and Osborne, 1970). Additionally, ethylene has been found to play a role in the response to both biotic and abiotic stresses, with its pathogen response functionality involving the synthesis of cell wall strengthening compounds, induction of several PR genes (Broekaert et al., 2006), and involvement in the synthesis of plant defensins — small, cysteine-rich peptides with predominantly antifungal properties (Penninckx et al., 1996).

Out of the hormones discussed here, ethylene signalling is the most complex. Ethylene receptors, located in the endoplasmic reticulum (ER), form a complex with CTR1, inducing the constant proteasomal degradation of EIN3 through binding by EBF1/2 and their associated proteins (Guo and Ecker, 2003). The presence of ethylene inactivates the negative regulation, leading to EIN3 being able to bind EBS (EIN3 binding sites) in the promoters of target genes, the most prominent of which include ERF transcription factors (Broekaert et al., 2006). The current hypothesis as to the exact details of the signalling involves EIN2, an ER protein constitutively phosphorylated by CTR1 preventing C-terminal cleavage and potentially marking EIN2 for proteasomal degradation. In the presence of ethylene, CTR1 is deactivated, resulting in EIN2 no longer being phosphorylated, leading to the C-terminal domain being cleaved off and relocated to the nucleus, where it inhibits the degradation of EIN3 (Ju et al., 2012).

1.2.5.5 Crosstalk between the signalling of different phytohormones

The signalling events mediated by each individual phytohormone outlined above are not independent of each other, and the triggered pathways connect and interact with each other (Pieterse et al., 2009). The best documented of these interactions is the antagonism between JA and SA, with JA conferring resistance to necrotrophic pathogens but susceptibility to biotrophic pathogens, and SA signalling leading to the opposite effect. A key player in this interaction is the transcription factor WRKY70, with the expression level of this convergent downstream target determining which of the two responses are going to be followed (Li et al., 2004). Overexpression of WRKY70 led to the induction of SA-induced PR genes, whilst repressing the JA marker gene PDF1.2, with suppression producing the inverse effects. Exposure of the altered expression lines to the biotrophic *Erysiphe cichoraceaurm* and necrotrophic *Alternaria brassiciola* produces the expected results, with the constitutive overexpressor being resistant to the biotroph and the antisense-silenced line being resistant to the necrotroph (Li et al., 2006). A number of other regulators of the antagonism have been identified at different stages of the pathways, with similar mutant line validation. MAP kinase 4 was found to be a highly upstream regulator, as it inhibits the action of EDS1 and PAD4, which lead to SA synthesis and the stifling of the JA pathway (Brodersen et al., 2006). NPR1, the mediator of the SA signal from the cytosol to the nucleus, was found to repress JA signalling whilst still in its cytosolic state (Spoel et al., 2003). The arrival of SA-induced NPR1 monomers in the nucleus activates GRX480, leading to a complex with TGA transcription factors that induces SA-responsive gene expression whilst repressing JA-responsive genes (Bari and Jones, 2009). MYC2, one of the transcription factors activated by the presence of JA, represses SA-induced gene expression (Laurie-Berry et al., 2006). It is quite plausible that this is the point of coronatine-induced susceptibility, as currently it is the only known SA signalling suppression event downstream of coronatine introduction (Pieterse et al., 2009).

The role of ethylene in the phytohormone crosstalk could best be approximated as an enhancer and a fine tuner. There is a great degree of connectivity and synergy between JA and ethylene signalling. A prime example of this comes in PDF1.2, which requires joint action of JA and ethylene signalling to be induced (Penninckx et al., 1998). JA signalling branches off into two separate paths based on the presence or absence of ethylene — MYC2 activation represses JA/ethylene genes such as PDF1.2 and promotes the expression of other JA response genes such as VSP2 and LOX2 (Lorenzo et al., 2004). In the presence of an ethylene stimulus, JA activity shifts away from MYC2 and towards ORA59 (Pré et al., 2008) and ERF1

(Lorenzo et al., 2003). This fine-tuning of the signalling is carried out through EIL1 and EIN3, which are bound by members of the JAZ family, which subsequently recruit HDA6 for repression through histone deacetylation. Upon exposure to both JA and ethylene, the JAZ proteins are degraded in a COI1 dependent manner and the two transcription factors become active, inducing downstream targets such as ORA59, PDF1.2 and ERF1 (Zhu et al., 2011). Somewhat more upstream in the pathway, ethylene has been proposed to modulate the NPR1-based antagonism point between SA and JA signalling (Leon-Reyes et al., 2009). Ethylene is also capable of enhancing the SA response, augmenting the expression of PR genes in an EIN2 dependent manner, suggesting a level of crosstalk between the cores of the two pathways (De Vos et al., 2006).

Given its natural tendency to induce heavy disease phenotypes even in cases of incompatible pathogen exposure, it should be of no surprise that the predominant functionality of ABA in the phytohormone crosstalk is the disabling of the defence response initiated by the other phytohormones. ABA has been shown to act antagonistically to JA/ethylene signalling (Anderson et al., 2004), as well as SA-induced SAR (Yasuda et al., 2008). However, ABA plays a vital role in the response to a number of abiotic stimuli, including drought. Therefore, ABA's ability to shut down the defence-related functionality of the other pathways likely serves as the determinant between the plant response to biotic or abiotic stimuli (Anderson et al., 2004). In some select cases, ABA can be involved in the fine-tuning of pathogen response by the core of the crosstalk. MYC2, which is the main transcription factor of the non-ethylene-stimulated branch of JA signalling, is known to be a positive regulator of ABA response (Abe et al., 2003), with the best known functionality of that branch of JA signalling being wound response (Anderson et al., 2004). When also factoring in ABA's ability to confer resistance to certain necrotrophic pathogens (Adie et al., 2007), ABA's highly specialised and not fully understood role in the biotic response crosstalk, likely to be primarily related to JA as shown by its involvement in the MYC2 branch in the wounding response, becomes apparent.

1.3 Transcription regulation and gene regulatory networks

As mentioned in the previous section, the plant response to both biotic (Windram et al., 2012) and abiotic (Kilian et al., 2007) stimuli involves large-scale transcriptional reprogramming. This allows the plant to alter the state of the cell and take appropriate action against the adverse condition affecting it. The transcriptional re-

programming response can be very swift — *A. thaliana* stimulated with flg22 showed differential expression of over 1000 genes in under an hour (Zipfel et al., 2004). The transcriptional overhaul is fine-tuned through a number of external stimuli, and the reprogramming itself is often carried out through a cascade of transcriptional regulation. For example, the jasmonic acid induced transcriptional response depends on the presence of other phytohormones — if other signalling is absent, the MYC2-dependent wounding response will be initiated, inducing the expression of transcription factors ANAC019, ANAC055 and ANAC072, which in turn reprogram the transcription of a number of downstream genes (Kazan and Manners, 2013). In the presence of ethylene, the transcriptional reprogramming shifts towards a pathogen defence response, with ERF1, PDF1.2 and ORA59 being the key downstream genes induced by EIL1/EIN3 (Zhu et al., 2011). It should come as no surprise then that the transcriptional regulatory events in plant cells take on the form of complex networks (webs of interactions between transcription factors and the induction of downstream targets), with the potential of crosstalk between networks responsible for responses to different stimuli (Figure 1.2, Shinozaki et al. (2003)).

1.3.1 General overview of transcription

Transcription is the process in which RNA polymerase produces mRNA reflecting the sequence of a gene stored within the cell's DNA. The mRNA is subsequently used as a template to synthesise the corresponding protein in a process known as translation. In conjunction, transcription and translation make up the cornerstone of what is known as the central dogma of molecular biology (Crick, 1970).

In eukaryotes, mRNA transcription is carried out by RNA polymerase II, which is a highly complex enzyme with a number of unique domains (Cramer et al., 2001). The polymerase forms a complex with a number of general transcription factors (GTFs), which carry out supplementary functions that help initiate and perform the process of transcription. The first of these to bind the DNA is transcription factor II D (TFIID), which recognizes the promoter (Orphanides et al., 1996). The promoter is the DNA sequence upstream of the gene coding region, designed to recruit the polymerase. It often features the TATA box, which helps attract the TATA binding protein (TBP) subunits of TFIID and indicate the direction of transcription with its asymmetry (Xu et al., 1991). In the absence of a TATA box, a promoter can still be recognised and used to initiate transcription upon detection of other sequences by TBP related factors (TRFs) (Hochheimer and Tjian, 2003). Upon TFIID binding, two further GTFs, transcription factor II A (TFIIA) and transcription factor II B (TFIIB), join the complex and promote the

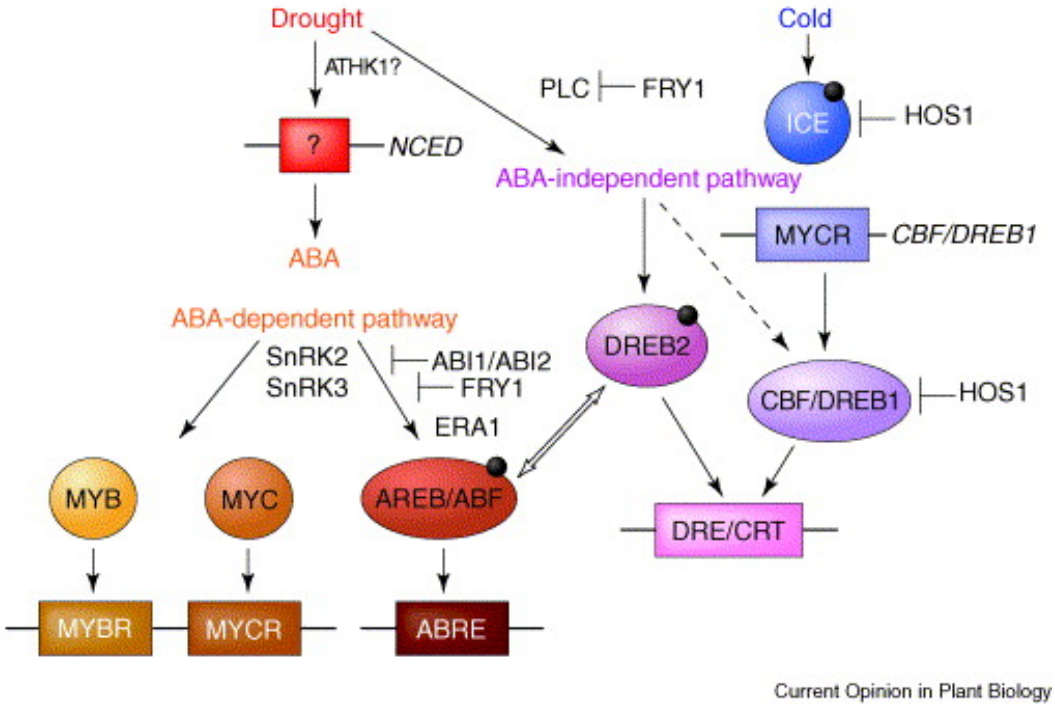

Figure 1.2: Crosstalk between regulatory networks in *A. thaliana* response to drought and cold. The schematic depicts the regulatory signalling triggered by the stimuli, with round nodes being transcription factors and the rectangular nodes being their corresponding binding motifs in the promoters of the genes they target. A part of the network is shared between both responses. Reproduced from Shinozaki et al. (2003).

binding and stabilisation of the remaining components of the transcriptional machinery, including the polymerase. The final GTFs to bind are transcription factor II E (TFIIE) and transcription factor II H (TFIIH), and once they are present transcription can commence upon ATP hydrolysis (Orphanides et al., 1996). TFIIH has been shown to be essential for the separation of DNA strands, creating what is known as the Open Complex, making it possible for the complex to synthesise mRNA complementary to the information stored within the DNA (Hahn, 2004).

In order to prioritise the transcription of relevant genes, the polymerase and its associated GTFs are guided to appropriate genes by transcription factors. Transcription factor proteins contain a number of easily discernible domains, with their functionality including the binding of specific DNA sites in the promoters of their regulated genes and being able to confer activation or repression to the transcription complex (Luscombe et al., 2000). The exact methods by which this is accomplished are quite varied, and include the summoning of chromatin modifying proteins to the transcription factor to make the DNA of the gene of interest accessible (Ikeda et al., 1999), binding of various components of the transcription machinery, most prominently TAFs, which are subunits of TFIID (Kornberg, 1999) and stimulating the rate of transcriptional initiation and elongation (Brown et al., 1998). Repression strategies include the dissociation of TBPs from the TATA box (Auble et al., 1997), binding to the activation domain of positive regulators (Ma and Ptashne, 1987), competing for the same DNA site (Vincent and Struhl, 1992) or promoting histone deacetylation, making the gene's coding region and promoter inaccessible (Grunstein, 1997). A number of these functionalities are carried out with assistance from modular protein complexes known as transcriptional co-activators (Näär et al., 2001). An example would be the complex that leads to transcriptional repression carried out by plant transcription factors with the EAR motif – SAP18 binds to the EAR motif on a transcription factor and in turn binds HDA19, which carries out histone deacetylation (Kagale and Rozwadowski, 2011).

An organism typically features a high number of transcription factors with a great deal of differing functionality – *A. thaliana* is predicted to have almost 2500 unique transcription factors spanning almost 100 superfamilies, with the ones with the highest number of members being MYB (147) and MYB-related (72), bHLH (149), C2H2 (148), AP2-EREBP (146), ND (145), NAC (111), MADS (106), HB (91), bZIP (75) and WRKY (74) (Pruneda-Paz et al., 2014). The superfamilies often feature internal structuring, with an example being the presence of the TGA family involved in salicylic acid signalling (Després et al., 2000), as well as AREB/ABF family involved in abscisic acid signalling (Fujita et al., 2013), in the bZIP superfam-

ily. The specificity of regulation mediated by a particular transcription factor stems from the characteristic interaction between the transcription factor's DNA binding domain and a binding sequence, specifically recognised by that domain, present in the target gene's promoter. An example of a characteristic binding sequence are ABRE (ABA-responsive elements) present in the promoters of genes regulated by the members of the bZIP superfamily involved in ABA signalling (Uno et al., 2000). Great advancements in the state of knowledge of transcription factor binding specificity have been made recently with the aid of protein binding microarrays, with this approach allowing for the experimental identification of binding motifs specific to individual transcription factors (Franco-Zorrilla et al., 2014; Pruneda-Paz et al., 2014). The specificity of these transcription factor-binding site interactions are the foundation for regulation and make it possible for fine-tuneable regulatory networks capable of inducing large-scale downstream expression overhaul to exist.

1.3.2 Searching for a network footprint

Initiating a transcriptional response results in the expression reprogramming of a high number of genes – for example, a time course of *A. thaliana* response to *B. cinerea* infection showed 9,838 genes become differentially expressed during the duration of the experiment (Windram et al., 2012). The response is a highly complex phenomenon, driving the expression change of individual genes in a multitude of different directions via elements of the regulatory crosstalk. Nevertheless, it is possible to identify the presence of an underlying regulatory phenomenon by identifying groups of putatively co-regulated genes from transcriptional data. Whilst it is impossible to elucidate co-regulation from transcriptional data with full certainty, a long-standing assumption deemed 'guilt by association' states that co-expression can be indicative of co-regulation (Altman and Raychaudhuri, 2001), and it has been shown that increasing the amount of expression data available in the analysis increases the rate of said assumption being true (Yeung et al., 2004). The identification of these downstream co-expression/co-regulation events, which could be called the footprint of the regulatory network, can be carried out via a multitude of computational approaches. Those can be divided into two primary method classes – clustering, which requires identified co-expression groups to exhibit similar behaviour across all of the analysed data, and biclustering, which can identify gene groups exhibiting similarity across subsets of the provided data. These approaches are further detailed in Figure 1.3 and the following sections.

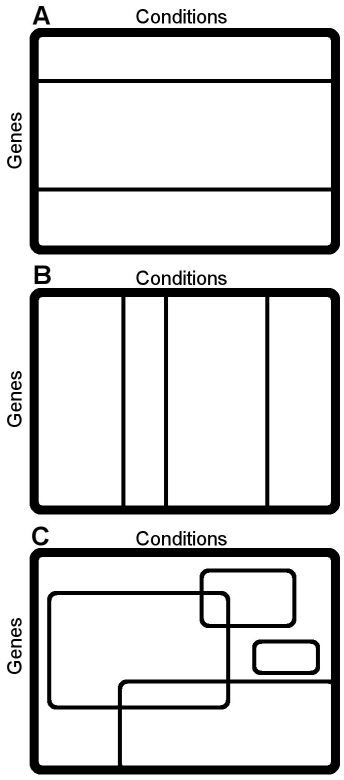

Figure 1.3: Comparison of clustering and biclustering. In the case of clustering approaches, the data is partitioned into groups, with all of the cluster members required to exhibit uniform behaviour across the entirety of the analysed data. This can be done on a per-gene **(A)** or per-condition **(B)** basis. Biclustering methods are capable of identifying subsets of genes and conditions where the co-expression occurs, as shown in **(C)**.

1.3.2.1 Clustering

The primary means for identification of groups of co-expressed genes is known as clustering, and its primary objective can be characterised as partitioning the input into distinct groups based on the provided data (Jain, 2010). It is an example of exploratory data analysis, wherein no pre-determined hypotheses or models exist and the data is mined for characteristic differentiating profiles (Tukey, 1977). The two primary kinds of clustering are hierarchical and partitional. Hierarchical clustering algorithms create a deterministic hierarchical structure from the data, starting out with each data point in its own separate cluster and proceeding to iteratively merge the most similar pair of clusters based on a similarity measure (such as Euclidean distance in an n-dimensional space made up of the observations for each data point) until it merges everything into a single cluster. Obtaining an actual cluster list involves placing a cut-off somewhere in the hierarchy. By contrast, partitional methods identify their clusters simultaneously and don't create a hierarchical structure of the data (Jain, 2010). An example of a partitional clustering algorithm is K-means, which places K potential cluster centres (either user defined or randomly generated) in the n-dimensional observation space along with

the data points, then assign each data point to the cluster centre nearest to it, then re-compute the cluster centre coordinates as the mean of the coordinates of all the data points assigned to it. The assigning data points to clusters and re-computing of the centre coordinates is repeated until convergence (Jain and Dubes, 1988). The procedure is not deterministic, in contrast to hierarchical clustering, but produces more evenly sized clusters. One of the problems with K-means is the identification of the optimal number of clusters, but heuristic methods combatting this issue have been proposed (Tibshirani et al., 2001).

Clustering has a firm role in the analysis of expression data and identifying potentially co-regulated genes. Application of clustering algorithms to expression data, be it temporal or a collection of multiple static measurements, reveals genes sharing functionality clustering together. This suggests a common regulatory mechanism driving the expression of genes with shared functionality, and implies the role of hitherto uncharacterised co-clustered genes in the same process (Eisen et al., 1998). An example of a clustering method capable of working with time series expression datasets is SplineCluster (Heard et al., 2005). The algorithm starts by fitting nonlinear splines to the expression profiles of individual genes, and proceeds to cluster the resulting fits in a Bayesian hierarchical framework. SplineCluster can also analyse multiple time courses at once — an application of the algorithm to multiple time courses of expression data from *Anopheles gambiae* challenged with *Escherichia coli*, *Salmonella typhimurium*, *Micrococcus luteus*, *Listeria monocytogenes* and zymosan revealed a number of gene clusters tightly co-expressed across all the stimuli. The identified clusters were evidently functionally enriched, further indicating co-regulation (Heard et al., 2005).

1.3.2.2 Biclustering

Whilst the impact of clustering on the ability to decipher regulatory interactions in expression data is undeniable, the inherent limitations of clustering algorithm design make the methods limited in the scope of research questions they can answer. In spite of the ability to manually curate the data used as input for the clustering, it is possible that the forced partitioning of the data will result in some uninformative genes being captured in the clusters. In addition, due to the complex and specific regulatory response to varying stimuli, it is possible that a shared response will only occur across a subset of the available data. Due to these constraints, an alternate class of methods termed biclustering were developed.

Representing the data as a matrix, with genes as the rows and observations as the columns, biclustering can be described as simultaneous clustering of the rows

and columns of this matrix, returning subsets of genes co-expressed across subsets of the available conditions as a result (Eren et al., 2013). Cheng and Church were the first to apply biclustering to expression data by proposing a deterministic greedy algorithm to mine the expression matrix (Cheng and Church, 2000). The method starts with the bicluster defined as the entirety of the matrix and progressively removes rows and columns from the bicluster to drop the mean square residue (MSR) below a given threshold. If multiple biclusters are to be found in the data, the discovered bicluster is masked with random values and the steps are repeated, making the method unfit for identifying overlapping biclusters (Cheng and Church, 2000). Another early biclustering approach is the Iterative Signature Algorithm (ISA), which creates an initial bicluster by taking a random sample of rows spanning all the columns, then proceeds to refine the row and column selection until obtaining sufficient value homogeneity (Bergmann et al., 2003). A known weakness of this method is its heavy predilection towards strong signals (Supper et al., 2007).

Since then, a wide scope of other biclustering methods have been developed, with a number of the algorithms having their design tailored to be able to answer specific research questions, including analyses of time course data (Eren et al., 2013; Tchagang et al., 2010). An example of a method designed for the analysis of multiple static expression data sets, ENIGMA accounts for the differential expression of the genes by examining the pattern of differential expression of gene pairs and translating those into a co-expression network, which is subsequently mined for modules (Maere et al., 2008). Enrichment Constrained Time-Dependent Iterative Signature Algorithm (ECTDISA) is a modification of ISA (Bergmann et al., 2003) for application to a single time course dataset by sampling single genes, constructing an initial module of the most correlated expression profiles and refining said module using a moving temporal window, with subsequent evaluation of the identified modules for biological functionality via automated scoring (Meng et al., 2009). It is also possible to include non-expression information directly into the biclustering, with cMonkey accounting for the presence of potential transcription factor binding sites in the promoters of genes whilst mining the expression dataset (Reiss et al., 2006).

A method of particular relevance is the Extended Dimension Iterative Signature Algorithm (EDISA), which is an extension of the ISA (Bergmann et al., 2003) algorithm to mine multiple time course datasets for genes co-expressed across a subset of the available conditions (Supper et al., 2007). This is of particular biological relevance, as temporal data provides a good degree of insight into the dynamics of the response and the crosstalk does not have to span all of the analysed datasets (Kilian et al., 2007; Kreps et al., 2002). Due to the shared ancestry with ECTDISA,

EDISA similarly starts by sampling a single gene from the dataset and creates an initial module around that gene by identifying the top co-expressed genes with it in one of the conditions. The module is then assumed to span across all of the conditions, and genes and conditions are iteratively removed until the module is either empty or classified as coherent (high Pearson Correlation Coefficient of expression profiles across all of the conditions jointly), independent (high Pearson Correlation Coefficient in each of the conditions separately, with no homogeneity of profiles required between conditions) or single response (one condition evaluated for fulfilling module conditions). The analysis finishes upon sampling a user-specified number of genes (Supper et al., 2007).

A drawback of EDISA and most other biclustering methods comes in the lack of evaluation of the identified co-expression patterns for actual regulatory roles. One of the most crucial advantages of having access to multiple datasets is the ability to tailor algorithms to discriminate between dependent co-expression, indicative of co-regulation, and independent co-expression occurring by pure chance due to the number of similar expression profiles and not reflecting an underlying shared regulatory mechanism. As an example, imagine 20 genes exhibiting similar behaviour across two time course datasets. If there were thousands of genes exhibiting that behaviour in each of the datasets separately, the overlap of 20 can be purely random and have nothing to do with a shared regulatory process, but a biclustering algorithm is likely to identify and report it. By contrast, if only 50 or so genes exhibited that behaviour in each of the datasets separately, the overlap of 20 becomes considerably more meaningful and hints at a shared regulatory interaction. This is accounted for by CCC-Biclustering, which discretises the expression profiles of genes into up-regulation, down-regulation or no change between every pair of time points, and subsequently assesses the chance that the regulatory phenomenon observed within a detected bicluster could have occurred by chance with the aid of the binomial distribution (Madeira et al., 2010), but it only operates on one time course and no method assesses this when using multiple time series in conjunction.

Clustering and biclustering algorithms lead to the detection of gene groups exhibiting co-expression across one or more expression datasets, with the co-expression having the potential to reflect co-regulation in spite of the lack of such assessment. It should be noted that no attempt is typically made to identify the putative regulators driving the identified co-regulatory events, and such inference typically requires mining for overrepresented transcription factor binding sites in gene promoters carried out through tools such as MEME-LaB (Brown et al., 2013).

1.3.3 Searching for gene regulatory network models

As mentioned when initially outlining clustering and biclustering algorithms, the results of such analyses are typically the downstream footprint of a series of regulatory interactions. Given an appropriate experimental design, it is possible to reconstruct these relationships and produce what is known as a regulatory network, which can potentially explain the circumstances leading to the observed co-expression of gene groups detected through clustering or biclustering. The term 'regulatory network' encompasses a very broad spectrum of possible models. Regulatory networks can be directly experimentally derived or indirectly inferred from experimental quantification, and they can capture interactions between proteins or the effects of transcription factors on gene expression (Windram et al., 2014).

1.3.3.1 Undirected co-expression networks

Regulatory networks can be represented in the form of a graph, with the graph's nodes being genes, and the edges capturing regulatory interactions between them. The graph can be undirected, with the edges capturing pairs of genes exhibiting similar expression profiles across the data used to infer the model, or directed, with the edges capturing precise hypotheses about one gene regulating another (Windram et al., 2014). Undirected regulatory networks, termed co-expression networks, are typically inferred from large collections of static expression data mined for pairs of genes that exhibit a high degree of correlation (typically evaluated with the aid of Pearson's Correlation Coefficient or Spearman's Correlation Coefficient) in expression across all of the available data (Usadel et al., 2009).

An example of this approach can be found in the inference performed on four separate datasets analysing citrus response to *Candidatus Liberibacter asiaticus*. Correlations for each pair of genes were computed for each of the datasets separately, resulting in individual correlation matrices. Those were subsequently combined into a single final correlation matrix by taking weighted sums of the individual matrices, with the weights reflecting the number of samples in each individual dataset. This resulted in a very large network structure upon the application of a stringent correlation threshold. In an attempt to elucidate functionality out of the massive model, a sub-network of genes with functionality related to SA signalling was identified (Zheng and Zhao, 2013). It is also possible to extend the methodology applied to the datasets beyond the basic correlation methods. An example would be the use of a graphical Gaussian model (GGM) (Toh and Horimoto, 2002). GGM is capable of accounting for the effect of other genes on the expression of a gene

pair by using partial correlations. As a practical example, if genes A and B are both regulated by gene C, they are likely to exhibit a high degree of similarity in their expression trends. By removing the influence of C on A and B, GGM reveals that there's no direct correlation between the pair. This method was applied to an extensive collection of static *A. thaliana* expression datasets, revealing a sparse underlying network structure predicting the crucial regulatory role of a few hub genes connected to a high number of other network nodes (Ma et al., 2007).

It is also possible to use the co-expression component in a broader approach, integrating data from other sources, such as protein-protein interactions, to create more complex model structures. An example is AraNet, which integrated a high number of genomic and proteomic datasets spanning multiple organisms into a massive functional network model with over a million connections of 24 distinct types, which led to the prediction of functionality for over 5000 *A. thaliana* genes with previously unknown roles, with subsequent experiments validating two of the three hypotheses tested (Lee et al., 2010). This modelling approach was subsequently fine-tuned and repeated with the addition of new datasets published since the initial network release, resulting in a sparser network (approximately 100,000 fewer connections across over 3000 more gene nodes) with predictive power for *A. thaliana* expanded through ortholog pairs into over 25 non-model plant species (Lee et al., 2014).

1.3.3.2 Ordinary differential equation models

Whilst informative in their own way, co-expression networks do not offer much information in the way of identifying explicit regulatory links between genes in the model. Alternate approaches have to be taken to produce a directed graph of a network, with the edges implying causality between the nodes. These are primarily inferred from time course data (Penfold and Wild, 2011). One of the methods that can be applied to produce such a result is modelling using a set of ordinary differential equations (ODEs). ODE systems are capable of accounting for a multitude of regulatory phenomena, featuring nodes in its network representing different forms of a gene, such as its mRNA, protein form and any complexes it may form. It is possible to capture the varying regulatory interactions happening at different layers, and produce an extremely detailed model. An example of ODE application is the inference of the *A. thaliana* circadian clock model (Pokhilko et al., 2010). The high amount of information available on the interactions and regulation of *A. thaliana* clock components allowed for the proposal of a system of 28 differential equations, with 43 parameters constrained based on prior knowledge and the remaining 61

inferred from multiple time course datasets. The resulting network model accurately captured the regulatory interactions between its components, and allowed for functional inference in the case of perturbations. An analysis of the network identified TOC1 as a repressor of LHY and CCA1, instead of an activator as previously believed, with the claim being subsequently experimentally validated with the aid of *toc1* mutant plants (Pokhilko et al., 2012). However, due to the high degree of prior knowledge required to construct an accurate model, ODE systems are best left to well characterised systems such as the *A. thaliana* clock, with other algorithms being more suitable for *de novo* network inference from large scale transcriptional datasets (Penfold and Wild, 2011).

1.3.3.3 Mining large-scale transcriptomic datasets for gene regulatory networks

Whilst it is possible to perform *de novo* inference of a regulatory network from a time course dataset featuring many genes, this procedure is likely to require a degree of pre-processing to be effective and computationally tractable. Common practice includes the identification of differentially expressed genes, reducing the gene list to transcription factors only in order to elucidate the core of the regulatory network, or performing clustering to group together genes with similar expression profiles (Penfold and Wild, 2011). Another limitation of the modelling is the sole reliance on mRNA levels, which are easily experimentally captured via microarrays (Schulze and Downward, 2001) or RNA-Seq (Wang et al., 2009), with the ability to experimentally quantify protein levels being limited in comparison (Miranda et al., 2007).

One of the classes of models that can be used to infer regulatory networks from large-scale transcriptomic data is dynamic Bayesian networks, with an example of an algorithm being Variational Bayesian State Space Modelling (VBSSM) (Beal et al., 2005). The underlying model structure of state space modelling (SSM) assumes the existence of a set of hidden states that drive the values of observed variables. For the purpose of this application, the observed variables are the expression profiles of the measured genes, whilst the hidden states capture all the effects that are missing from the experimental quantification, such as unmeasured genes or mRNA degradation. The basic SSM is expanded to include an additional driving input to affect the hidden state change between time points, and said input is set to be the gene expression at the previous time point. A simple transformation of the resulting equations allows for the identification of a single matrix capturing the interactions between every possible gene pair in the data, where interaction can

be defined as transcriptional activation or repression based on transcript levels in adjacent time points of the parent (regulating) and child (regulated) gene. This matrix is iteratively elucidated from the data in a process named the Variational Bayesian Expectation Maximisation (EM) Algorithm, which is a modification of the EM algorithm that accounts for the parameters forming a distribution instead of a point estimate. Subsequent iterations aim to maximise the marginal likelihood, and the final model can be converted into a binary network by applying a z-score threshold to the final parameter matrix (Beal et al., 2005).

Another possible modelling approach is Causal Structure Inference (CSI) (Klemm, 2008; Penfold and Wild, 2011). Similarly to VBSSM in nature, it infers its connections as a relation between the expression of the parent at a given time point and the expression of the child at the succeeding time point. The algorithm relies on Gaussian processes to accomplish this task — given n possible parents in a given evaluated fit, an n+1-dimensional space is created, with n of the axes occupied by the time shifted parents and one by the child, and a zero-mean Gaussian process prior is fit to the data. The hyperparameters of the Gaussian process fits are tuned with the EM algorithm to maximise the marginal likelihood, defined as the sum of the likelihoods of all of the individual fits performed as part of CSI. In order to combat the scaling of the number of fits in need of evaluating to preserve computational tractability, the concept of indegrees is introduced, which caps the maximum number of parents to be evaluated for the child in the fit. Upon completing the fits, each individual fit can be compared with the others by the computed marginal likelihood. The likelihoods are pooled together and scaled to produce a marginal distribution, which allows the extraction of the frequency with which any particular parent appears across the models (Penfold and Wild, 2011).

Given the existence of numerous network inference algorithms (other options include methods using Granger's causality), a comparison of the accuracy of network reconstruction of the individual algorithms was performed (Penfold and Wild, 2011). The process was performed using DREAM4 data, which is a collection of synthetic data reflecting the topology of known regulatory networks in *Escherichia coli* and *Saccharomyces cerevisiae*. The data collection features ten different networks, with five each featuring 10 and 100 genes, and the expression data being 21 time points modelled using parameterised stochastic differential equations (Greenfield et al., 2010). The performance of the algorithms was assessed using the area under the ROC curve, capturing the changes of true and false positive rates across varying stringency thresholds, and the area under the precision-recall curve, which captures the changes of true positive rate and positive predictive value in a similar

fashion. For the 10-gene networks, VBSSM outperformed all other options, including dynamic Bayesian networks with no hidden states, for three of the five networks, with CSI achieving the best results for the other two. For the 100-gene networks, CSI had the best scores for every network model tested (Penfold and Wild, 2011).

An example of the application of such methodology to large-scale transcriptomic data can be found in the analysis performed on a high resolution (24 time point) time course of *A. thaliana* response to *B. cinerea* infection (Windram et al., 2012). The full data set features 30,336 microarray probes, with this number being far beyond the computational tractability scope of any network inference approach. The first step in decreasing the dimensionality of the data was the manually curated identification of differentially expressed genes with the aid of GP2S (Stegle et al., 2010), which reduced the number of genes to 9,838. Whilst this is a step in the right direction, the number of profiles was still too high for network inference algorithms to handle. As such, SplineCluster (Heard et al., 2005) was applied to the set of differentially expressed genes, resulting in 44 co-expressed gene clusters, which was sufficient to perform network inference. The clusters, along with a tubulin expression profile representing *B. cinerea growth*, were mined for an underlying network model with CSI (Klemm, 2008; Penfold and Wild, 2011). The resulting network structure was used to formulate a number of specific regulatory hypotheses by matching transcription factors from specific families placed directly upstream of clusters with known binding sites for that transcription factor family overrepresented in the genes' promoters. The model also helped cement the novel role of TGA3 in necrotrophic pathogen response, subsequently experimentally validated with the aid of mutant lines, by placing its cluster (where it was one of only two transcription factors) as the only node upstream of the *B. cinerea* growth profile in the network model (Windram et al., 2012).

1.3.3.4 Experimentally derived networks

The creation of regulatory network models does not have to rely only on computational inference. A number of dedicated experimental approaches, focusing on the elucidation of instances of transcription factors binding DNA sequences, can be used to directly identify potential regulatory interactions taking place within the organism, detecting the paths that regulatory signal flow can take in response to a stimulus. Depending on the exact nature of the posed biological query, different techniques can be applied. In the case of a single promoter query of interest, which is to be scanned for potential binding by multiple transcription factors, yeast one-hybrid (Y1H) can be utilised. Y1H features the introduction of two plasmid

constructs to *S. cerevisiae*, with one of them featuring a fragment of the promoter sequence of the downstream gene of interest followed by a reporter gene, such as luciferase, and the other containing a fusion of the putative upstream regulator of interest with a strong transcriptional activation domain, such as VP16. If the regulator binds the promoter fragment, transcription of the reporter gene is induced (Ouwerkerk and Meijer, 2001). A number of libraries, featuring *S. cerevisiae* strains transformed with appropriately modified transcription factors, have been created for increased analysis throughput, with an example *A. thaliana* resource spanning 1956 transcription factors (Pruneda-Paz et al., 2014). However, the method suffers due to being performed in *S. cerevisiae* instead of *in planta*, which makes it difficult to reconstruct more intricate or condition-specific regulatory information. If a single regulator is of interest, and its role genome-wide is to be assessed, ChIP-seq is preferable. The proteins bound to the sample DNA are fixed by subjecting to formaldehyde, and the DNA with the bound proteins is isolated from the sample. The chromatin is subsequently sonicated to break it up into fragments, with fragments bound by the transcription factor of interest subsequently isolated with the aid of a specific antibody. The transcription factor-bound fragments are subsequently sequenced, and can be later mapped to the genome, identifying genes that the transcription factor regulates by binding to their promoters. An immediate improvement of this method over Y1H comes in its *in planta* nature, allowing for the capturing of actual regulatory events happening within the organism, but the execution requires antibodies rigorously tested for binding specificity (Park, 2009).

As previously stated in section 1.3.3.1, network modelling can take on an integrative form, with a large quantity of data from many different sources coming together to create a model. Experimentally derived transcription factor-DNA interactions can form the core of such models, with an example being a network formed from Y1H-interactions interactions of a number of transcription factors with the promoters of ANAC019, ANAC055 and ANAC072. The initial core was then refined on a per-condition basis through the application of hCSI and dedicated mutant line testing (Hickman et al., 2013). Due to such integrative possibilities, some experimental techniques are explicitly designed for use with other, potentially computational approaches. An example are protein binding microarrays, which are capable of determining the specific DNA sequences that a given transcription factor binds. This is accomplished through the exposure of a single transcription factor to a comprehensive collection of all 10-mer double stranded DNA sequences, with the bound sequences subsequently detected through tagging the attached transcription factor with fluorescent antibodies (Berger and Bulyk, 2009). The resulting binding

motif information can be used with computational approaches to propose downstream targets for the examined transcription factor in a regulatory network.

1.4 Aims and organisation of this thesis

The ability to identify a network footprint in a set of time course expression datasets is a very important part of trying to unravel the shared regulatory mechanisms between the stimuli, as it provides insight into the functionality of network regulation. At the start of my research, no clustering or biclustering approach was capable of mining multiple time course datasets for modules of genes co-expressed across a subset of the available conditions whilst accounting for whether the identified co-expression patterns are identified by chance or indicative of a shared regulatory process. Computational inference of the underlying network model can follow, but due to the complexity and tractability of the analysis these models have to be limited in scope, resorting to a subset of the available genes or representative clusters.

In this thesis, attempts are made to address some of the current shortcomings of co-regulatory data mining and network inference. The development of new algorithms, as well as application and expansion of previously existing methodology, allows for the computational inference of easily experimentally testable regulatory hypotheses, which in turn can lead to broadening the understanding of plant stress responses. The work was carried out using a number of available *A. thaliana* time course datasets showing response to a number of both biotic (Windram et al., 2012) and abiotic (Breeze et al., 2011) environmental conditions. The thesis consists of a published manuscript, a submitted manuscript currently under review and a manuscript that will be submitted shortly.

In Chapter 2, Wigwams identifies genes working across multiple situations (Wigwams) is introduced. Wigwams is an exploratory data mining algorithm that performs a complete search for dependently co-expressed gene modules spanning all possible multi-condition subsets of the provided time course datasets, and evaluates the co-expression it detects for statistical significance with the aid of a modification of the hypergeometric test. The method is applied to a selection of both biotic and abiotic *A. thaliana* stress response time course datasets, and uncovers a number of modules with overrepresented functionality among their members with known transcription factor binding motifs in their promoters, implying the means through which the relevant functionality is jointly regulated across condition-specific responses.

In Chapter 3, Wigwams is applied as part of the analysis of three separate time courses capturing the response of *A. thaliana* to two strains of *P. syringae*:

the virulent wild type *P. syringae* pv. *tomato* DC3000 and the avirulent hrpA mutant that is unable to deliver effectors to the plant. Wigwams uncovers a number of modules showcasing varied response profiles to both the virulent and avirulent pathogen strain, and indicates a shared ABA signalling mechanism with rice as well as uniform down-regulation of chloroplast-related genes in response to the detection of both pathogens. Other analyses performed on the data include the elucidation of a time scale of pathogen response events, using both single time point differential expression analysis and a Gaussian process gradient tool application to delta profiles, defined as the log ratio of two of the time courses, revealing a very abrupt early PTI spike and subsequent effector-mediated down-regulation of assorted chromatin stability genes around 6 hours post infection. A joint CSI model for early response transcription factors across all three time courses was also inferred, with an analysis of the hub genes driving the highest number of network nodes with extreme differences in hrpA and DC3000 response implying a defence role of XND1 and novel candidate genes FBH3 and AT2G40620.

In Chapter 4, Wigwams is applied to a number of high-resolution time course datasets showcasing *A. thaliana* response to a number of biotic and abiotic stimuli, revealing a number of regulatory footprints shared by up to five of the six tested conditions. A number of the findings are consistent with the results of the analyses performed in Chapter 3, including the identification of multi-condition down-regulation of nucleosome and chloroplast genes. Wigwams is also applied to large-scale network models inferred via a modification to VBSSM, utilising its statistically significant co-expression indicative of co-regulation to expand the transcription factor-only models by matching module membership to network edges. The resulting enhanced network models have their nodes functionally annotated by assessing functionality overrepresentation of their immediate downstream targets, and a more detailed analysis of the *B. cinerea* and *P. syringae* networks reveals a number of both novel and previously known genes annotated with GO terms related to the four primary defence hormones (JA, SA, ABA, ethylene), leading to the identification of a five-gene interaction which may play a role in the signalling mediated by all four of those hormones in both conditions.

In Chapter 5, the significance, merits and limitations of the methods and approaches utilised in Chapters 2 through 4 are discussed, along with outlining the developments in experimentally derived knowledge of transcription factor binding sites and computational tractability of network inference algorithms evident in work conducted at different stages of the project. A number of ideas with regards to future improvements to the computational inference of gene regulatory inference are

proposed, including potential extensions to the introduced methodology, application of further experimental datasets, subject to availability, and the need for optimal algorithm implementation.

Chapter 2

Wigwams: identifying gene modules co-regulated across multiple biological conditions

The work featured within this chapter has been published in the journal Bioinformatics (Polanski et al., 2014). I am joint first author. All supplementary data for this chapter can be accessed as part of the article on the website of the journal: `http://bioinformatics.oxfordjournals.org/content/30/7/962.full`

A sound biological question is the identification of genes exhibiting co-regulation across multiple time course expression experiments, indicating a likely shared regulatory process controlling their expression under those stimuli, whilst at the same time not requiring the identified co-expression to span across all of the provided datasets or forcibly partitioning the data. For example, a group of genes may be co-regulated during fungal pathogen infection and senescence, but this regulatory mechanism may not span bacterial pathogen infection. The available algorithms capable of mining multiple time course datasets for modules of genes co-expressed across subsets of the available conditions, such as EDISA (Supper et al., 2007) and tensor methods (Li et al., 2012; Zhang et al., 2012), do not attempt to assess whether the co-expression they detect is indicative of co-regulation. Additionally, these methods do not account for the differential expression status of the genes across conditions, and might identify genes not involved in the response to particular conditions as part of their modules. This motivation led to the creation of Wigwams, which is capable of identifying modules of co-expressed genes spanning subsets of multiple time course datasets, whilst assessing whether the detected co-expression is indicative of co-regulation and accounting for the differential expression status of

genes in each individual condition.

Dr Johanna Rhodes proposed the original implementation of Wigwams in her PhD thesis (Rhodes, 2012). She introduced the discrimination between dependent and independent co-expression through the use of the hypergeometric test, a statistical approach used for the evaluation of the significance of overrepresentation, with bioinformatics applications such as GO term analysis (Maere et al., 2005). The hypergeometric test allowed for co-regulation assessment by computing the chance of the observed overlap between the top genes co-expressed with a single reference gene in two different time series occurring by chance. In the case of a lack of a shared regulatory mechanism across the evaluated conditions, the observed overlap should be low, giving no grounds to reject the null hypothesis of no co-regulation. In the case of a shared regulatory event across the conditions, the detected overlap should be higher, leading to the rejection of the null hypothesis and identification of a potentially co-regulated gene module. In order to identify all such modules contained within the data, a complete search was performed, with each gene being used as the reference gene for module creation, resulting in a thorough but redundant module list showing genes potentially co-regulated across pairs of conditions. The redundancy was subsequently removed in a procedure known as pruning, where modules for each pair of conditions were sorted on p-value and filtering was performed for each condition combination separately. This allowed modules with smaller p-values to remove modules with inferior p-values if they were sufficiently similar in gene content (and as such redundant). Afterwards, subsetting reconstructed modules spanning three or more conditions out of the subset pair-of-condition modules if they identified similar genes as co-regulated, giving reason to believe that the co-regulation extended beyond two conditions.

The original Wigwams implementation outlined above had a number of elements in need of optimisation — the original redundancy removal and multiple condition module reconstruction approach led to a loss of information (measured as unique genes featured in the modules) whilst still leaving a number of uninformative modules in the final output. Crucially, the co-regulation evaluation was only performed for pairs of conditions. Formulating and implementing a generalisation of the pairwise testing procedure resulted in a drastic increase of the share of modules spanning three and more conditions in the module total. This obsoleted subsetting, but created the need for dealing with redundancy among modules spanning subsets of the same conditions. Unique gene loss was greatly decreased by merging redundant modules instead of removing them, allowing the algorithm to reconstruct larger scale regulatory events. The original implementation was also imperfect in

places, with the reimplementation removing a number of bugs and greatly increasing computational efficiency.

For this study, I redesigned the algorithm as described, conducted the analyses, interpreted the results, wrote the manuscript, compiled the supplementary data, and made figures under the guidance of Dr Katherine Denby, Dr Sascha Ott and Dr Bärbel Finkenstädt.

Wigwams: identifying gene modules co-regulated across multiple biological conditions

Krzysztof Polanski[1,†], Johanna Rhodes[1,†,‡], Claire Hill[2], Peijun Zhang[2], Dafyd J. Jenkins[1], Steven J. Kiddle[1,§], Aleksey Jironkin[1], Jim Beynon[1,2], Vicky Buchanan-Wollaston[1,2], Sascha Ott[1] and Katherine J. Denby[1,2,*]

[1]Warwick Systems Biology Centre and [2]School of Life Sciences, University of Warwick, CV4 7AL, UK

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[‡]Present address: NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Foundation Trust, London SE5 8AF, UK.
[§]Present address: Department of Infectious Disease Epidemiology, Imperial College, London W2 1PG, UK.

Motivation: Identification of modules of co-regulated genes is a crucial first step towards dissecting the regulatory circuitry underlying biological processes. Co-regulated genes are likely to reveal themselves by showing tight co-expression, e.g. high correlation of expression profiles across multiple time series datasets. However, numbers of up- or downregulated genes are often large, making it difficult to discriminate between dependent co-expression resulting from co-regulation and independent co-expression. Furthermore, modules of co-regulated genes may only show tight co-expression across a subset of the time series, i.e. show condition-dependent regulation.

Results: Wigwams is a simple and efficient method to identify gene modules showing evidence for co-regulation in multiple time series of gene expression data. Wigwams analyzes similarities of gene expression patterns within each time series (condition) and directly tests the dependence or independence of these across different conditions. The expression pattern of each gene in each subset of conditions is tested statistically as a potential signature of a condition-dependent regulatory mechanism regulating multiple genes. Wigwams does not require particular time points and can process datasets that are on different time scales. Differential expression relative to control conditions can be taken into account. The output is succinct and non-redundant, enabling gene network reconstruction to be focused on those gene modules and combinations of conditions that show evidence for shared regulatory mechanisms. Wigwams was run using six Arabidopsis time series expression datasets, producing a set of biologically significant modules spanning different combinations of conditions.

Availability and implementation: A Matlab implementation of Wigwams, complete with graphical user interfaces and documentation, is available at: warwick.ac.uk/wigwams.

Contact: k.j.denby@warwick.ac.uk

Supplementary Data: Supplementary data are available at Bioinformatics online.

2.1 Introduction

Elucidating the regulatory mechanisms mediating biological processes is a key challenge in many eukaryotic organisms. Much regulation occurs at the transcriptional level; however, despite our ability to profile genome-wide gene expression and the availability of bioinformatics tools to analyze sequence information, our understanding of gene regulatory networks underlying biological processes is still relatively basic. Regulatory interactions are often common, meaning that the ability to understand the regulation of a response requires a mathematical or computational network model. Underlying these network models is the knowledge of regulatory mechanisms. Techniques to identify regulatory mechanisms, such as genome-wide chromatin immunoprecipitation sequencing (Robertson et al., 2007) and matrix-based yeast one-hybrid (Y1H) (Deplancke et al., 2006), have improved, but these techniques are not high-throughput. Therefore, it is crucial to be able to make high-quality predictions of regulatory mechanisms using existing data; these predictions can then be tested in focused experimental and modelling efforts.

Time series experiments are often used to examine the dynamics of gene expression (Belling et al., 2013; Windram et al., 2012), and with the decreasing cost of profiling techniques, more datasets covering multiple time series showing how an organism responds to different conditions are becoming available. Such data offer the opportunity to study shared regulatory mechanisms that are used to regulate genes under more than one condition. Such shared regulatory mechanisms may drive gene expression in the same way in each time series, or they may be modified to drive expression on a different time scale or to change the direction of regulation (activation versus repression). To facilitate gene network reconstruction, it is important to develop tools that can map out which gene modules are co-regulated in what combinations of conditions.

It is a long-standing assumption that co-expression may reflect co-regulation (Altman and Raychaudhuri, 2001), and using data from multiple conditions can improve the correlation between the two (Yeung et al., 2004). However, in noisy biological systems, co-regulated genes may still show some differences in their expression, and there may be more than one regulatory mechanism that can drive genes with a particular expression pattern. Following a perturbation, such as infection or treatment with a chemical stimulus, many genes change in expression and even with high-resolution time series, large numbers of genes can show a similar expression profile (Weinstock-Guttman et al., 2003; Windram et al., 2012). As a result, it can be challenging to distinguish between *dependent co-expression* indica-

tive of co-regulation and *independent co-expression* of genes that achieve a similar expression pattern in different ways. It is important to note that both dependent and independent co-expression may pass statistical tests that are geared towards testing the similarity of expression patterns and/or the tightness of a gene cluster relative to other clusters. Therefore, tools that solely aim to detect clusters of similarly expressed genes may not discriminate informative dependent co-expression from uninformative independent co-expression.

Multiple high-resolution time series of gene expression for a single organism under different conditions provide a powerful approach for identifying dependent co-expression of genes likely to be controlled by a common upstream regulator. However, while an increasing number of time series would help to improve the specificity of co-expression (i.e. co-expression across more time series is more likely to be dependent co-expression), it is unlikely that a single group of genes will be co-expressed across all the datasets. For example, it is known that there is significant cross-talk between signalling networks for different plant hormones in Arabidopsis, but not all of the components are playing a role in the response to every hormone (Robert-Seilaniantz et al., 2011). Furthermore, some of the detected co-expression across multiple datasets may still be independent co-expression due to the abundance of particular expression profiles rather than due to a shared regulatory mechanism. Hence, there is a need for an algorithm that can identify modules of genes dependently co-expressed across subsets of time series data, combining the increased specificity of multiple time series datasets with biological reality.

A myriad of clustering algorithms have been developed to assign genes into clusters based on the similarity of their expression profiles across a single time series or multiple static (i.e. single or few time points) datasets (Madeira et al., 2010; Maere et al., 2008; Meng et al., 2009; Reiss et al., 2006). A few algorithms have also been developed to specifically cluster genes using two or more time series datasets, such as BHC (Savage et al., 2009) and SplineCluster (Heard et al., 2005). However, these algorithms partition genes into clusters and do not enable identification of genes co-expressed across subsets of the data. A few methods are capable of identifying co-expression across subsets of the data, but these come with their own drawbacks. Tensor methods (Li et al., 2012; Zhang et al., 2012) require the timescale of the experiments to be uniform throughout. EDISA (Supper et al., 2007) does not require the same timescale across all of the datasets, but it is non-deterministic. None of these methods is able to incorporate differential expression relative to control time series into the analysis, and crucially, none statistically evaluates dependent co-expression versus independent co-expression. ENIGMA (Maere

et al., 2008) can account for genes' differential expression, but the method was designed for a series of static expression data. CCC-Biclustering (Madeira et al., 2010) tests biclusters for statistical significance against a null hypothesis of independent expression profile evolution, but the method is only capable of analyzing a single time course experiment.

Wigwams (*Wigwams identifies genes working across multiple situations*) is a simple, deterministic and efficient method capable of identifying groups of dependently co-expressed genes, termed gene modules, spanning subsets of the available time series datasets. Wigwams is a comprehensive method; all potential dataset combinations are scanned for gene modules by rigorously testing putative gene modules around each gene differentially expressed in a dataset. Wigwams evaluates each putative module for statistical significance and provides a non-redundant output of gene modules showing significant co-expression across varying combinations of the time series data. Each gene may be assigned to one or more gene modules or to no module at all. Wigwams requires little user input (further aided by easy-to-use graphical user interfaces) and is computationally inexpensive and relatively fast, making it a useful method to analyze multiple time series experiments for evidence of co-regulation. We demonstrate that gene modules identified by Wigwams are often enriched for Gene Ontology (GO) terms (Ashburner et al., 2000) and known transcription factor (TF) binding motifs indicating biological relevance. We also provide experimental evidence of potential co-regulation. Wigwams is a powerful tool to utilize the resolution of time series expression data in a statistically rigorous approach for identification of co-regulated gene modules. It can make a direct contribution to gene regulatory network analysis and computational prediction of regulatory mechanisms.

2.2 Materials and methods

Here we outline the Wigwams algorithm indicating the various steps and decisions to be taken in applying this method to multiple time series expression datasets. A Matlab implementation is provided at warwick.ac.uk/wigwams along with relevant documentation.

2.2.1 Input

The input to Wigwams is a matrix of gene expression values for all the time series data to be analyzed with unique gene identifiers and annotation of each time series sample. Two additional matrices can be provided to improve the biological relevance

and ease of interpretation of resulting modules: one indicating which genes are differentially expressed (DE) in each time series dataset (previously determined relative to a control time series) in a binary manner (0 for non-DE, 1 for DE), and the second providing annotation information for each unique gene identifier. If data on DE genes are not provided, then all genes are treated equally. A graphical user interface has been created to aid in the construction of data formats for use in Wigwams based on raw input files. A second graphical user interface facilitates running the individual steps of the Wigwams method described in Sections 2.2.2, 2.2.3 and 2.2.4 below.

The expression profiles are standardized on a per-gene basis in each dataset, and a matrix containing the expression profiles of all genes differentially expressed in at least one of the conditions is created. The expression profiles of non-DE genes within each condition are randomly shuffled across non-DE genes. This prevents non-DE genes from contributing to gene modules, as it unlinks dependencies of expression profiles across conditions for those genes (see Section 2.2.2 below). Any non-DE gene making a (coincidental) contribution to a gene module is removed from the module (see below). Therefore, although the randomization step eases the data processing, it has no effect on the eventual output of Wigwams, leaving the Wigwams output deterministic.

2.2.2 Identifying modules spanning multiple datasets

This stage of Wigwams is outlined in Supplementary Figure S1, with an example shown in Figure 2.1. The aim of these steps is to detect all evidence of co-regulation (in the form of dependent co-expression) across the multiple time series datasets, regardless of the redundancy of resulting modules. Each gene that is DE in two or more conditions is deemed a 'seed' gene and is tested sequentially. For each condition in which the seed gene is DE, the other genes in the expression matrix are ranked on the basis of how well their expression profile in that time series is correlated with the expression profile of the seed gene. Genes are ranked with the most correlated gene at the top of the list. In the work presented in this article, the Pearson correlation coefficient was used as the similarity measure. Alternative metrics could be substituted without a need to change the Wigwams method itself.

For each combination of conditions in which the seed gene is DE, the size of the overlap between the top-ranked co-expressed genes in each dataset is tested statistically (Fig. 2.1 shows an example). This evaluates whether the similarities of gene expression observed within each time series are dependent across datasets. A significant P-value suggests that a regulatory mechanism is at work that (i) targets a

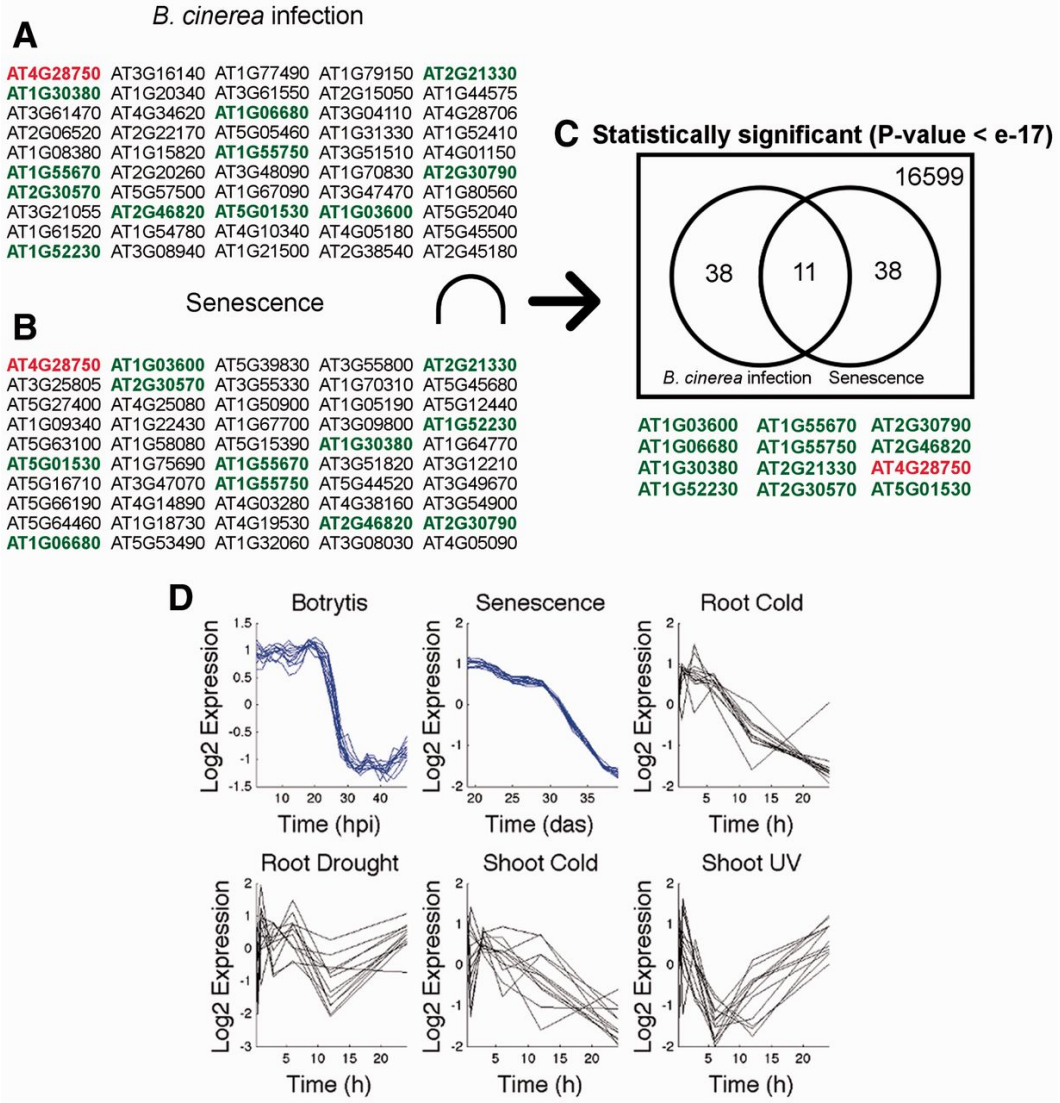

Figure 2.1: Strong evidence for dependent co-expression is detected during the module identification stage. **(A)** and **(B)** are two lists of 50 genes that are most correlated to the seed gene's expression profile in *B.cinerea* infection and senescence, respectively (seed gene shown in red). **(C)** The overlap between the two lists (genes shown in green) is determined, and the expression profiles of the identified genes are shown in **(D)**. To evaluate whether the observed overlap is likely to have occurred by chance a hypergeometric test is performed (yielding a P-value below 1 e-17 in this example). Overlaps deemed statistically significant are likely to discriminate dependently co-expressed genes from independently co-expressed genes. Such overlaps are therefore termed 'modules'. In this example, the module is 'spanning' *B.cinerea* infection and senescence. Hpi, hours post inoculation; das, days after sowing, h, hours

significantly similar set of genes in each condition considered and (ii) induces expression profile similarity in each condition. However, no restriction is made regarding the similarity of gene expression profiles across different conditions. Therefore, a regulatory mechanism that targets a similar set of genes under different conditions, but with a different effect on expression (e.g. activating in one condition but repressing in another) can still be detected by this statistical test.

All possible combinations of conditions are tested (i.e. sets of two or more time series up to the number of independent time series used), using the top n most correlated genes in each dataset, where the user specifies a range of values for n. In our work, we have used $n = 50$, 100, 150, 200 and 250 to be able to detect regulatory mechanisms targeting < 10 to > 100 genes. Wigwams processes the detection of modules (described in this Section) independently for each n and pools the results.

To determine whether the observed overlap is statistically significant, the hypergeometric test is used. For the purposes of the Matlab implementation of Wigwams, the hypergeometric test function by Meng et al. (2009) was used. This was modified to enable the significance of overlaps between more than two time series to be assessed. Given a universe of size U and two sets of size $N_2 = [n_1, n_2]$, the probability of observing an overlap of at least size x by chance equals

$$H_2(N_2, x, U) = \sum_{i=x}^{\min(N_2)} \frac{\binom{n_1}{i}\binom{U-n_1}{n_2-i}}{\binom{U}{n_2}}$$

One can expand this for k sets of size $N_k = [n_1, n_2, \ldots, n_{k-1}, n_k]$ by assuming the probability of observing an overlap of at least size x between the k sets to be equal to the sum of the products of the probability of observing an overlap of exactly size i between $k-1$ sets of size $N_{k-1} = [n_1, n_2, \ldots, n_{k-2}, n_{k-1}]$ and the probability of observing an overlap at least of size x between two sets of size $[i, n_k]$, for $x \leq i \leq \min(N_{k-1})$. Owing to the nature of the hypergeometric test, the probability of observing an overlap of exactly size i between $k-1$ sets is equal to the difference of the probability of observing an overlap at least of size i and the probability of observing an overlap of at least size $i+1$ for $i < \min(N_{k-1})$, and to $H_{k-1}(N_{k-1}, \min(N_{k-1}), U)$ for $i = \min(N_{k-1})$. Combining that into a formula yields

$$H_k(N_k, x, U) = H_{k-1}(N_{k-1}, \min(N_{k-1}), U) \cdot H_2([\min(N_{k-1}), n_k], x, U)$$
$$+ \sum_{i=x}^{\min(N_{k-1})-1} [(H_{k-1}(N_{k-1}, i, U) - H_{k-1}(N_{k-1}, i+1, U)) \cdot H_2([i, n_k], x, U)]$$

This modification makes it possible to evaluate the statistical significance of overlaps between three and more time series datasets.

The Bonferroni correction (Bland and Altman, 1995) is applied to the P-values from the hypergeometric test. Given a desired significance threshold α (0.05 was used for this study), the Bonferroni correction proposes an adjusted α

$$\alpha_{corr} = \alpha / \sum_{i=1}^{N} (2^{n_i} - n_i - 1)$$

where n_i is the number of datasets in which gene i is differentially expressed. N is the number of genes. Overlaps with a P-value below the adjusted significance threshold were deemed to have statistically significant dependent co-expression. Such overlap gene lists are considered gene modules and always include the seed gene by design. If any non-DE genes are included in these overlaps, these are removed from the putative gene modules before evaluating the statistical significance. Therefore, the output is a list of gene modules showing statistically significant dependent co-expression across two or more time series datasets. However, at this stage multiple modules may contain similar genes as if seed genes have similar expression profiles, similar gene modules will be created around these.

2.2.3 Merging similar modules spanning the same time series subset

This process in Wigwams is outlined in Supplementary Figure S2, with an example shown in Figure 2.2. It is important that the output of Wigwams is in a useful format for biologists to use, and hence at this stage of the algorithm, gene modules with a sizeable overlap of gene membership are merged. However, this stage only reduces redundancy among modules that are spanning the same combination of conditions.

Owing to the way modules are formed in Wigwams, modules with large overlap will also have similar expression profiles. In this study, modules with an overlap $> 30\%$ of the smaller module's gene membership are merged. This simplifies the

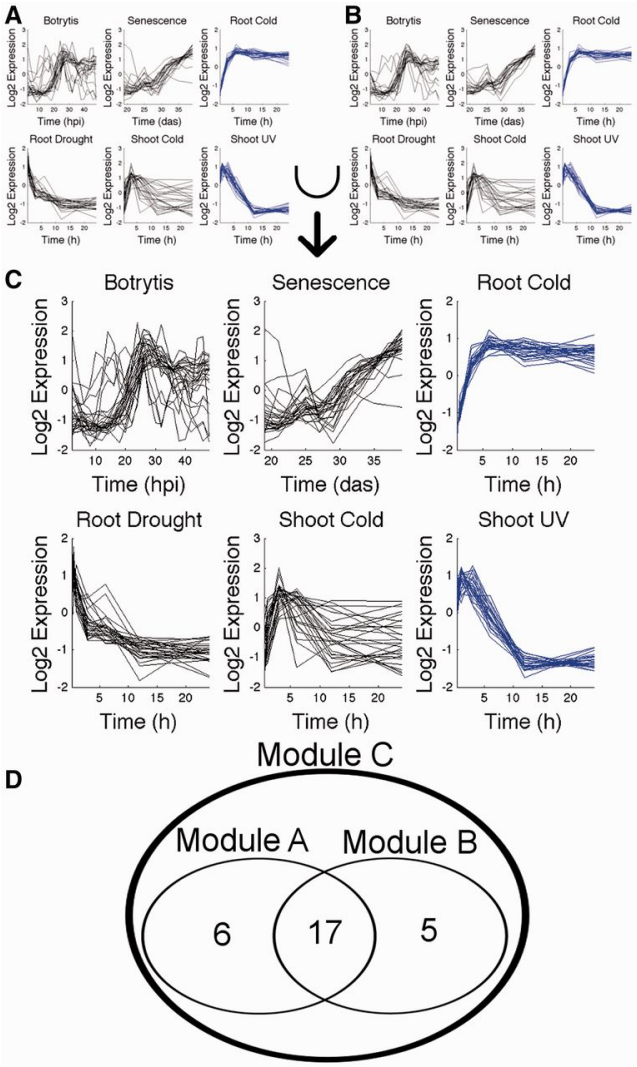

Figure 2.2: Merging. The modules shown in (**A** and **B**) span the same combination of conditions (depicted in blue; other time series shown in black), and feature a sizeable overlap in gene membership. Merging joins the two modules into a combined module shown in **(C)**. The mean expression profiles of the larger module A are used to determine whether the five genes unique to module B are expressed with sufficient similarity to be included in the joint module, preserving the extra information that is contributed by module B. In this example, all genes of module B are accepted. Hpi, hours post-inoculation; das, days after sowing, h, hours

Table 2.1: Gene module information during Wigwams analysis

Wigwams stage	Raw	After merging	After sweeping	After thresholding
Modules	4434	161	128	78
Overlaps	38964	4	3	3
Max overlap	50	19	19	19
Two conditions	3465	44	39	39
Three conditions	787	70	50	26
Four conditions	173	40	33	12
Five conditions	8	6	5	1
Six conditions	1	1	1	0
Mean module size	22	56	63	100
Total size	97313	9030	8050	7827
Unique genes	4444	4239	4197	4194

Note: The table shows the number of modules, number of pairs of modules that span the same condition combination with at least 10 genes in common (overlaps), maximum number of genes shared by a pair of modules spanning the same conditions (maximum overlap), number of modules spanning two to six conditions, mean module size, total size of the identified modules and number of unique genes they feature (Unique genes) for the initial module list (raw) and at different stages of Wigwams analysis.

output because one module, containing genes co-expressed with similar profiles, is formed rather than two. The choice of overlap threshold was based on the distribution of overlap size (as a proportion of the smaller module) as seen in Supplementary Figure S1. The distribution of overlap size can be plotted within Wigwams providing a tool for the user to set this threshold. Genes from the smaller module are added to the larger module provided that their expression profile is sufficiently correlated with the mean expression profile of the larger module. We used a correlation threshold of 0.8, which needed to be fulfilled for each of the datasets the modules span. In addition, two modules whose mean expression profiles across relevant time series datasets are highly correlated (Pearson correlation coefficient of at least 0.9 for each of the datasets) are also merged, regardless of the overlap in gene membership. The merging stage produces a set of modules with greatly reduced redundancy but without loss of essential information (see Table 2.1). Broader regulatory phenomena are reconstructed using previously identified statistically significant modules. The user can adjust the thresholds to shift the trade-off between the ability to see subtle differences between similar modules and the ability to get a succinct overview of key signals in the data.

2.2.4 Sweeping redundant modules spanning different dataset subsets

This stage of Wigwams is outlined in Supplementary Figure S3, with an example shown in Figure 2.3. Sweeping addresses a second kind of redundancy. For example, in the case of a module containing genes significantly co-expressed across three conditions, significant dependent co-expression is likely to be picked up for each pair of these time series as well, yielding another three modules. The gene membership of the module spanning more conditions is compared with that of modules spanning subsets of these time series. If the overlap is comparable with the size of the module spanning fewer conditions, then this module is discarded on the basis of it not contributing significant new information. In this study, the module spanning fewer conditions was discarded if the overlap featured at least 50% of its gene members.

The output at this final stage of Wigwams is a list of modules generated from genes showing statistically significant dependent co-expression, processed to optimize the number of different expression patterns contained in these modules and reduce redundancy between module gene membership.

2.2.5 GO term and TF binding motif enrichment testing

GO term (Ashburner et al., 2000) enrichment was tested with the Cytoscape plugin BiNGO (Maere et al., 2005) using the `GO_Full` ontology with the hypergeometric test and the Benjamini-Hochberg correction to control the false discovery rate (Benjamini and Hochberg, 1995). The whole genome Arabidopsis annotation was used as a reference set. Analysis of overrepresented TF binding motifs in module promoter sequences was carried out exactly as in Breeze et al. (2011), using information from the PLACE (Higo et al., 1999) and TRANSFAC (Matys et al., 2006) databases. P-values were adjusted using the Benjamini-Hochberg correction. For each gene, 500 bp of DNA upstream of the transcriptional start site was tested. As a control for GO term and TF binding motif analysis, 78 groups of genes were randomly generated from the 16 686 genes forming the Wigwams input. These 78 random modules were the same size as the 78 final modules identified by Wigwams.

2.2.6 Yeast one-hybrid technique

The yeast one-hybrid TF library screen was performed as described in Hickman et al. (2013) using three overlapping promoter fragments of ∼400 bp spanning ∼1 kb upstream of the transcription start site of each gene.

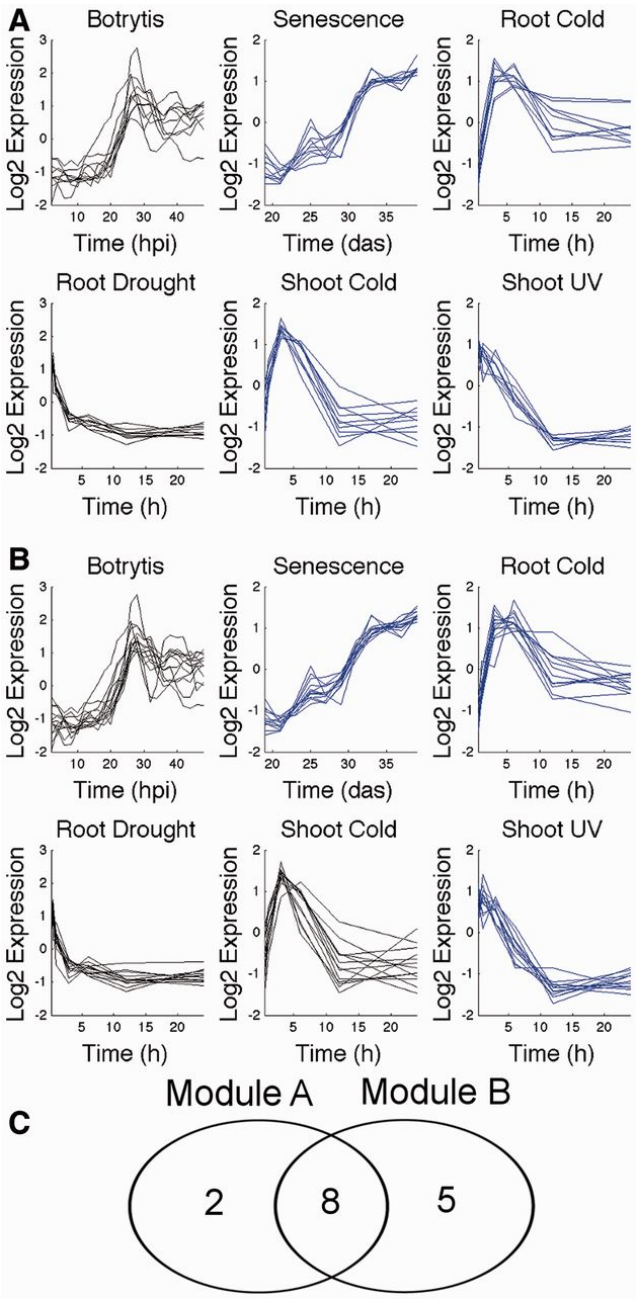

Figure 2.3: Sweeping. Sweeping evaluates those pairs of modules where one spans only a subset of conditions compared with the other. The module spanning fewer conditions is removed if it does not contribute enough new information. In this example, module A spans four conditions (senescence, root response to cold, shoot response to cold, shoot response to UV, shown in blue), while module B only spans three of these and contains only five genes that are not already members of module A. Module B is discarded. hpi, hours post-inoculation; das, days after sowing, h, hours

2.3 Results

We applied Wigwams to analyze a set of six time series datasets of gene expression responses to environmental stress in the model plant *Arabidopsis thaliana*. Two of the datasets were high-resolution time series, one obtained from leaves following infection with the fungus *Botrytis cinerea* (Windram et al., 2012), the other from leaves developing from maturity to senescence (ageing) (Breeze et al., 2011). The other four datasets had fewer time points, and captured responses to abiotic stresses [shoot and roots after cold stress, roots during drought and shoots after ultraviolet (UV) light exposure]. These were obtained from AtGenExpress (Kilian et al., 2007). The two groups of experiments were performed with different microarray platforms (Redman et al., 2004; Sclep et al., 2007), and the datasets were found to have 19 886 genes in common.

For the *B.cinerea* infection and senescence time series, the curated lists of DE genes were used (Breeze et al., 2011; Windram et al., 2012). For the AtGenExpress datasets, differential gene expression was determined using the GPTwoSample test (Stegle et al., 2010), with a score threshold of four. In all, 16 686 genes were DE over time in at least one condition (Supplementary Dataset S1), with 12 447 genes DE in at least two conditions and hence eligible for inclusion into Wigwams modules.

2.3.1 Wigwams systematically scans the data for evidence of co-regulation

The module identification procedure uses one gene at a time ('seed gene') and each combination of conditions in turn and tests whether the expression pattern of the seed gene across these time series may be driven by a regulatory mechanism acting on a number of genes under more than one condition. This is illustrated in Figure 2.1 for the case of a set of two conditions. For each time series, gene expression similarity to the seed gene is evaluated and the list of genes that are most strongly correlated with the seed gene assembled. Genes in each list show co-expression (across multiple conditions), which could be dependent co-expression driven by a common mechanism, or independent co-expression where multiple mechanisms induce similar expression patterns. If the expression similarity observed in each time series is the result of a common regulatory mechanism, then it is likely that this mechanism will target a similar set of genes in each condition. Wigwams tests this hypothesis. In the example of Figure 2.1, of 50 genes in each list, 11 genes (plus the seed gene) are in common between the two lists. By the hypergeometric test, the likelihood of making this observation by chance is below 1e-17. This provides

strong evidence that the co-expression observed is not independent co-expression, but dependent co-expression driven by a shared regulatory mechanism. Hence, the 11 genes in the overlap (plus the seed gene) are likely to be under a common regulatory influence and are considered a module.

The module identification procedure was run for all DE genes and all dataset combinations. This stage of the analysis took 2 h 53 min on a Dell Precision M4700 computer (2.8 GHz Intel Core i7-3840QM processor, 16 GB DDR3 SDRAM at 1600 MHz, 64-bit Windows 7 Professional, Matlab R2012b), producing a list of 4434 statistically significant gene modules likely to be showing dependent co-expression spanning two to six conditions (Table 2.1). Of the 12 447 DE genes in two or more conditions, 4444 were placed in at least one module.

2.3.2 Wigwams effectively removes redundancy among modules

As Wigwams considers every DE gene as a seed gene during the module identification stage, the method is comprehensive, but the output after the first stage is likely to have a high degree of redundancy. The merging algorithm merges modules with similar gene membership and/or highly similar expression profiles (exemplified in Fig. 2.2). The sweeping algorithm removes modules that have a large overlap with another module, but only show dependent co-expression across a smaller subset of conditions (exemplified in Fig. 2.3). In both cases, the essential information characterizing the expression phenomenon observed is maintained, while redundant information is removed.

After the merging stage, the initial 4434 modules were condensed into 161 modules (Table 2.1), while the number of unique genes assigned to at least one module only decreased from 4444 to 4239. The genes lost during merging had expression profiles not sufficiently similar to the mean expression profile of the larger module to be included. The average size of modules increased from 22 to 56 genes, while overlaps among modules were strongly reduced. Redundancy within the remaining 161 modules was further reduced by the sweeping stage. This reduced the list to 128 modules with an average size of 63, while the number of unique genes included in modules decreased only slightly from 4239 to 4197.

We decided to exclude small modules from further analyses, as (i) we wanted to get an overview of expression signatures driven by major regulatory mechanisms and (ii) although the excluded modules did pass rigorous statistical testing, the evidence base for these modules is not as wide as for the larger ones. We required a minimum of 10 genes for modules spanning two or three conditions, a minimum of 8 if spanning four, and a minimum of 5 genes if spanning five or six time series. These

thresholds were simply chosen on the basis that fewer genes will be co-expressed across a larger number of datasets and small modules will provide little functional information. Our thresholding resulted in a final list featuring 78 modules spanning two to five conditions and covering 4194 unique genes (Table 2.1 and Supplementary Dataset S2). The mean module size is 100 genes.

2.3.3 Wigwams reveals expression signatures of regulatory mechanisms

Four modules from the set of 78 are shown in Figure 2.4 (expression profiles for all modules are in Supplementary Dataset S3). Strong evidence for dependent co-expression has been detected for time series coloured in blue. The evidence for co-regulation of genes in these modules does not merely stem from the tightness of expression patterns, but is further supported by the dependence of expression similarities across time series. Although some expression profiles appear correlated in conditions not part of the module (e.g. shoot cold in Fig. 2.4D), we have not found evidence for dependence of co-expression in these time series. Expression similarity arises from a large number of genes sharing a similar profile in that condition.

Interestingly, module A shows regulation in different directions depending on conditions. During *B.cinerea* infection, senescence and root response to cold genes in the module are upregulated, while they are downregulated during root response to drought and shoot response to UV, consistent with the idea that the mechanism regulating these genes operates in a different mode under different conditions.

The full set of 78 modules contains modules dependently co-expressed across different combinations of conditions (Fig. 2.5 and Table 2.1). The modules detected by Wigwams can be considered the result of regulatory networks active during different stress responses. By analyzing the distribution and function of modules across different combinations of conditions, hypotheses can be made about the function and complexity of the networks underlying these, and network reconstruction attempts can be directed towards suitable time series combinations. For example, a researcher interested in unravelling shared regulatory networks between the biotic stress of *B.cinerea* infection and abiotic stress can see from Figure 2.5 that only one module spans *B.cinerea* infection, root response to drought and shoot response to cold. As such, attempts to reconstruct a common regulatory network spanning these three stress responses do not appear to be well supported by the available data, as we see little evidence for a complex shared network. In contrast, there are nine modules spanning *B.cinerea* infection, root response to cold and shoot response to UV light, making this combination of stresses much more promising for elucidating a

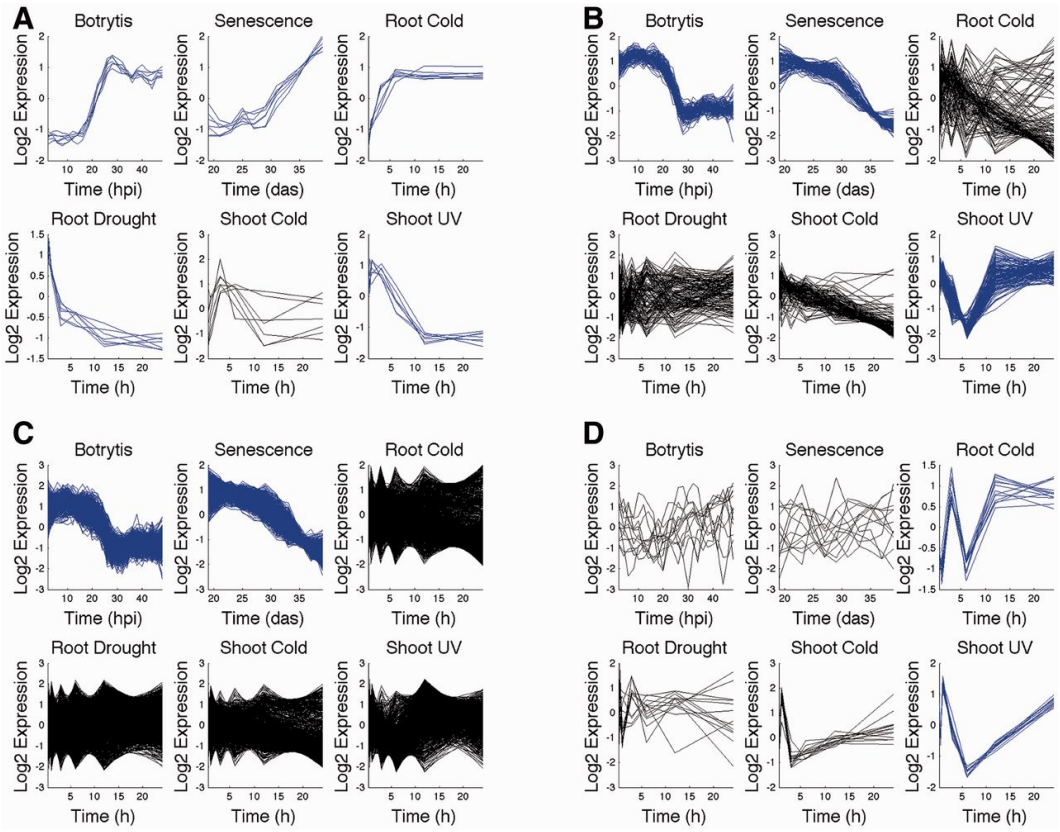

Figure 2.4: Four examples of modules showing different regulatory phenomena detected by Wigwams. Each module is represented by the gene expression profiles of its members across the six conditions. Shown in blue are conditions for which there is evidence for dependent co-expression. **(A)** is the smallest module, which has seven genes and appears to be dependently co-expressed in every condition except for shoot response to cold. The genes are activated in three conditions and repressed in two. **(B)** is a 131-gene module spanning *B.cinerea* infection, senescence and shoot response to UV. **(C)** is the largest module with 1238 genes, including all 131 genes from module **(B)**, but only spans two conditions. **(D)** features 13 genes with unusual expression profiles in root response to cold and shoot response to UV. hpi, hours post inoculation; das, days after sowing; h, hours

common regulatory network. Wigwams allows the researcher to rigorously examine available data for evidence of a regulatory network before embarking on modelling or experimental efforts.

2.3.4 Biological validation of detected modules

The enrichment of genes involved in the same biological process is often used as an indication of co-regulation of a gene module. Therefore, we tested Wigwams modules using BiNGO (Maere et al., 2005) and found that 71 of the 78 gene modules were overrepresented for GO terms (Ashburner et al., 2000) relating to biological processes vis-à-vis 24 of the 78 random modules. This further supports the case for co-regulation of genes in each Wigwams module. Two examples of such modules, along with the identified overrepresented GO terms, are shown in Figure 2.6 (overrepresented GO terms for all modules are given in Supplementary Dataset S4).

The module shown in Figure 2.6A features 29 genes spanning four datasets and is enriched in GO terms 'response to abscisic acid (ABA)' and 'response to cold'. This suggests a wider role of genes responding to cold and a role of ABA, a plant hormone, in mediating the link between the four conditions. The module shown in Figure 2.6B contains 269 genes dependently co-expressed across shoot response to cold and UV light, and is enriched in GO terms corresponding to the CUL4 RING ubiquitin ligase complex and non-coding RNA processing. The enrichment of genes with a shared function (members of the same complex) suggests that a specific mechanism acts to co-ordinate expression of the genes in this module.

Transcriptional gene regulation occurs by the binding of TFs to specific DNA sequences in promoters of genes. The same or similar TF binding motifs are often present in the promoters of co-regulated genes. We tested the Wigwams modules for enrichment of known TF binding motifs and found that 51 of the 78 modules had at least one overrepresented motif (Supplementary Dataset S5), suggesting that the module members were co-regulated. In comparison, 6 of the 78 random modules were overrepresented for a motif. The promoters of genes in the module shown in Figure 2.6A were overrepresented for the W box (de Pater et al., 1996), suggesting that the dependent co-expression is driven by binding of TFs from the WRKY TF family (Eulgem et al., 2000), known to play a key role in regulating plant stress responses (Chen et al., 2012b; Eulgem and Somssich, 2007). The module shown in Figure 2.6B was strongly enriched in a motif bound by the TCP TF family (Cubas et al., 1999; Welchen and Gonzalez, 2006), again suggesting a mechanism for shared regulation.

Finally, we tested the validity of the Wigwams modules using the Y1H tech-

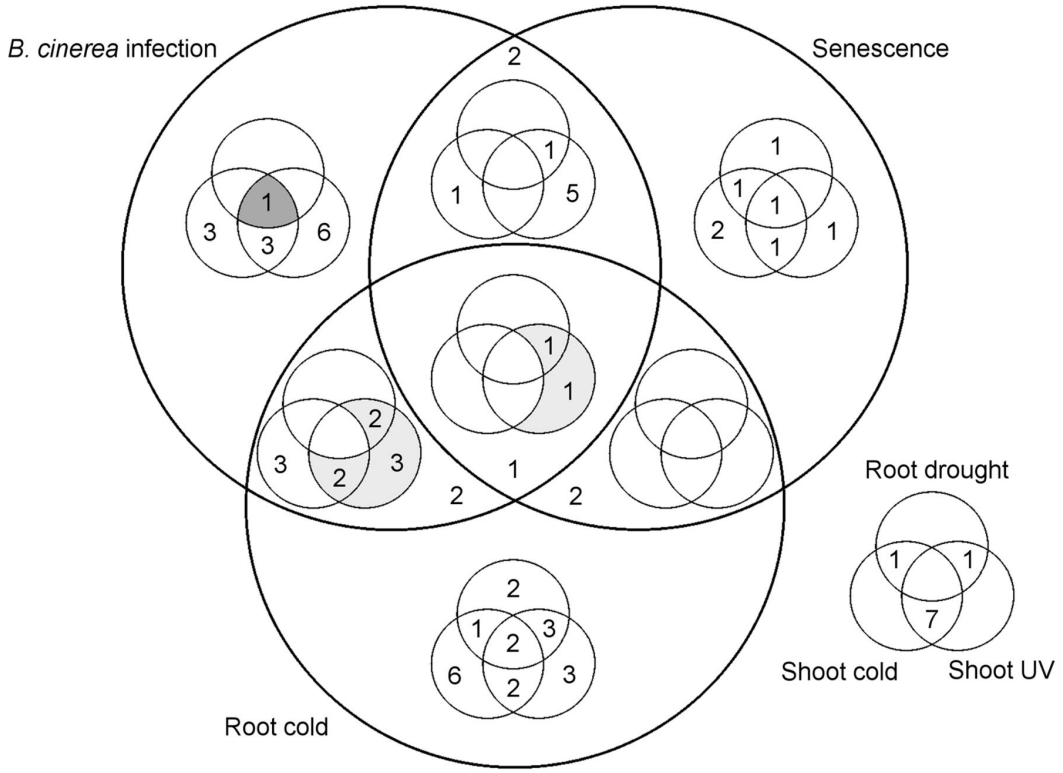

Figure 2.5: The number of modules identified for each combination of conditions. Three conditions are represented by large circles; the other three by small circles. Evidence for dependent co-expression was found across a range of combinations of conditions, ranging from two time series up to five time series. By analyzing the number of modules detected for different combinations of conditions, network reconstruction efforts can be focused on time series combinations showing evidence for shared regulatory mechanisms. The nine modules featuring *B.cinerea* infection, root response to cold and shoot response to UV light are shaded light grey. The single module spanning *B.cinerea* infection, root response to drought and shoot response to cold is shaded dark grey

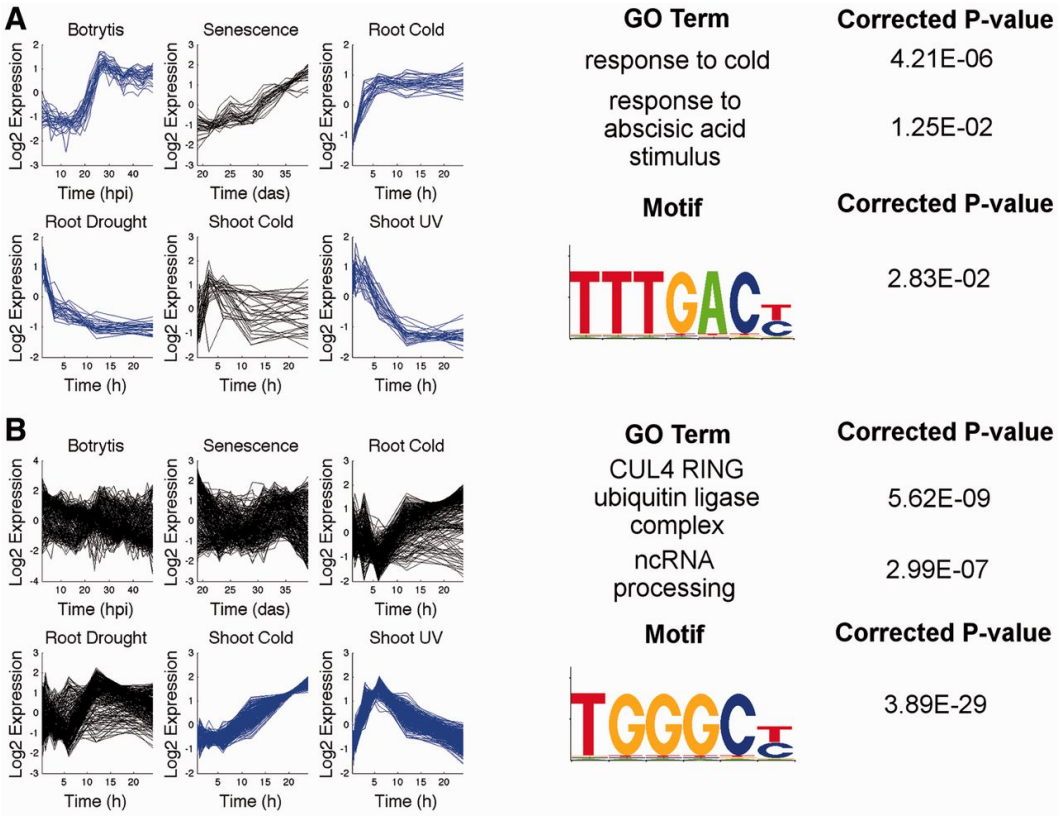

Figure 2.6: Wigwams modules are enriched in GO-terms and TF binding motifs. Shown above is an excerpt of the biological information obtained for two modules, showing dependent co-expression (in blue) and overrepresentation of GO terms and TF motifs in the promoters for genes in each module. **(A)** A 29-gene module spanning four conditions suggests a role for abscisic acid in linking the transcriptional responses to these four conditions. **(B)** A 269-gene module spanning shoot response to cold and UV light shows highly significant overrepresentation for the TCP binding motif, suggesting this motif may be underlying the dependent co-expression driving ubiquitination and non-coding RNA processing. hpi, hours post-inoculation; das, days after sowing, h, hours

nique to test for direct binding of the same TF(s) to multiple genes within a module. As we are interested in gene regulatory networks, we targeted TF gene promoters from Wigwams modules. The promoters of two genes (AT3G15210 and AT5G05410) from a 26-gene module spanning *B.cinerea* infection and root and shoot responses to cold (Supplementary Fig. S4a) were screened against a TF library to identify TFs able to bind these DNA sequences. Both of these promoters were bound by TCP3 (AT1G53230) and TCP1 (AT1G67260), two members of the TCP TF family (Martín-Trillo and Cubas, 2010). In a 38-gene module spanning senescence and the root response to drought (Supplementary Fig. S4b), the promoters of three genes (AT1G19180, AT1G80840 and AT3G23250) were screened and were bound by WRKY41, a member of the WRKY TF family (Eulgem et al., 2000). Direct binding of these TFs to multiple gene promoters from the same module is a strong indication that the Wigwams algorithm is detecting dependent co-expression reflecting co-regulation.

2.4 Discussion

Wigwams is a simple deterministic method capable of identifying groups of genes exhibiting statistically significant dependent co-expression across subsets of time series datasets, and using that information to construct larger non-redundant modules capturing broader transcriptional phenomena. Its comprehensive nature minimizes the odds of missing evidence of co-regulation, and the redundancy removal procedures provide the researcher with a succinct biologically informative output. In some cases, when examining the expression plots of gene modules, the module appears to exhibit co-expression in conditions that are not deemed significant. This demonstrates the power of Wigwams to select modules with statistically significant dependent co-expression. In these non-significant conditions, the given expression profile may have been abundant and/or the module members are not DE in that time series (i.e. expression profile in the control samples was similar).

When comparing Wigwams with other methods capable of identifying groups of genes co-expressed across different subsets of time series data, its main advantages are flexibility, statistical significance testing and relevance of the provided output. Additionally, Wigwams is able to account for differential expression of genes in each of the time series, and ensure that gene profiles are only tested for statistically significant dependent co-expression in relevant conditions. The value of testing the statistical significance of detected co-expression can be seen when comparing Wigwams with the EDISA algorithm (Supper et al., 2007). When run on a permuted

dataset, EDISA identified several co-expressed gene modules, whereas Wigwams did not identify any. Furthermore, we have shown the value of the comprehensive nature of Wigwams; it is capable of detecting dependent co-expression that EDISA misses (see Supplementary Material for details on this analysis). To our knowledge, Wigwams is the only algorithm capable of mining multiple time series (on varying time scales) for dependent co-expression across subsets of the time series.

The modules produced by Wigwams were demonstrated to be biologically relevant due to enrichment of GO terms (Ashburner et al., 2000) and known TF binding sites, suggesting shared function and regulation between module members. We also provide experimental evidence for co-regulation showing that in yeast, a set of similar TFs bind to the promoters of multiple genes from a single Wigwams module. Although Y1H does not indicate the conditions under which these TFs bind to the gene promoters, or whether they bind *in planta*, it does indicate the potential for co-regulation.

The Wigwams tool is easy to use, with intuitive graphical user interfaces, comprehensive documentation and output provided in a clear manner that can be readily analyzed by tools such as BiNGO (Maere et al., 2005) and MEME-LaB (Brown et al., 2013). The algorithm is flexible, and intuitive parameters can be used to tailor the output as desired. Additionally, the module lists are saved as Matlab cell structures, enabling access to intermediate stages of Wigwams analysis, e.g. to identify the most statistically significant original smaller gene modules.

A more computationally tractable version of the modified hypergeometric test could enable modification of the Wigwams method. To obtain the P-values for an overlap spanning k sets, all the P-values for $2, 3, \ldots, k-1$ sets need to be generated. If the tests were more efficient, the algorithm could be modified to use correlation thresholds instead of pre-defined set sizes when evaluating overlaps, and non-DE genes could be excluded from any analysis without large adverse effects on run time due to varying universe size between dataset combinations.

Owing to the time and cost of experimental approaches to genome-wide network elucidation, computational inference of regulatory networks from time series expression data is a useful approach. However, despite the multitude of inference methods available, these methods are still only capable of inferring 'moderately large dynamic networks' (Kim et al., 2013). Wigwams provides output that can be used to extend network models built with a subset of genes (e.g. using TFs only). Integrating Wigwams modules with a transcriptional network model can also provide condition-dependent information, such as indicating network neighbourhoods active during particular conditions. Wigwams modules can be viewed as the footprint

of flux through regulatory networks under different conditions, and examining the abundance and functionality of modules for various combinations of conditions can provide insight into the commonality between the responses to different conditions at a more nuanced level than simple differential expression. Identification of modules showing contradictory expression under different conditions (e.g. upregulated in one dataset and downregulated in another) also suggests points of cross-talk within the regulatory network.

Funding: K.P., S.J.K. and A.J. were funded for this work by the Engineering and Physical Sciences Research Council (EPSRC)/Biotechnology and Biological Sciences Research Council (BBSRC) funded Warwick Systems Biology Doctoral Training Centre; J.R. was funded by a BBSRC Systems Approaches to Biological Research studentship; C.H., D.J., P.Z., J.B., V.B-W., S.O. and K.J.D. are part of the BBRSC funded grant Plant Response to Environmental Stress Arabidopsis (BB/F005806/1). The TF library was a gift from Franziska Turck, Max Planck Institute, Cologne, Germany.

Conflict of Interest: none declared.

Chapter 3

Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in Arabidopsis leaves following infection with *Pseudomonas syringae* pv. *tomato* DC3000

The work within this chapter has been submitted to the Plant Cell journal. I am joint first author. The version of the manuscript in this thesis is a revised manuscript re-submitted on September 18th 2015. Once published, the supplementary data will be available online on the journal's site, and can currently be accessed at `http://www2.warwick.ac.uk/fac/sci/sbdtc/people/students/2011/krzysztof_polanski/thesis_supplement/` (password: 21datasets).

The motivation behind this work was to perform a highly precise deconstruction of the *Arabidopsis thaliana* defence response to *Pseudomonas syringae* pv. *tomato* DC3000 infection. A key feature unique to this study is the replication of the experiment for an avirulent strain of the pathogen with a key component of the se-

cretion system knocked out, which renders it unable to deliver effectors to the plant. This allows the time course data to capture the transcriptome-wide mechanics of uninterrupted PTI and contrast them against the effector-altered expression in the leaves infected with the virulent strain. Mining the data involved the initial identification of differentially expressed genes, followed by a reconstruction of temporal dynamics of regulatory events based on their times of first differential expression. A follow-up analysis examining the behaviour of genes across the different treatments revealed effectors from the virulent strain suppressing a number of PTI response processes, including the repression of abscisic acid-responsive genes by infection with the avirulent strain. The effectors were not merely limited to altering the expression of genes differentially expressed between the mock and avirulent strain, but also induced and repressed distinct sets of genes unaltered during PTI. These included genes involved in ubiquitination (upregulated by effectors) and chromatin assembly (downregulated). Gene grouping (performed through both clustering and Wigwams) was subsequently used in conjunction with transcription factor binding site analysis to identify potential regulatory interactions both specific to individual treatments and conserved across the whole experiment. A transcription factor-only regulatory network model was inferred across all of the experiments, identifying a number of both known and novel genes as potential defence response hubs.

My involvement with this study, under the guidance of Dr Katherine Denby, Prof. Murray Grant and Dr Sascha Ott, was as follows:

- Conducted the analyses, interpreted the results, wrote the manuscript, compiled the supplementary data, and made figures for the work shown in Figures 3.6, 3.7, 3.8, S6, and supplementary data sets 6, 9, 10, 11, 12
- Cleaned up contradicting data left by Siddharth Jayaraman, used it to compile supplementary data set 1 and the scatterplots in Figure 3.1
- Performed the Gaussian process gradient tool analysis used as the foundation for Figures 3.2, 3.3 and S3
- Extracted the relevant ubiquitination and chromatin-related genes, created the heatmaps in Figure 3.4 and compiled supplementary data set 5

Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in Arabidopsis leaves following infection with *Pseudomonas syringae* pv. *tomato* DC3000

Laura A. Lewis[1,2,4], Krzysztof Polanski[1,4], Marta de Torres-Zabala[3], Siddharth Jayaraman[3,φ], Laura Bowden[α], Jonathan Moore[1], Christopher A. Penfold[1], Dafyd J. Jenkins[1], Claire Hill[2], Laura Baxter[1], Satish Kulasekaran[3], William Truman[3,#], George Littlejohn[3], Justyna Prusinska[2], Andrew Mead[2,§], Jens Steinbrenner[2], Richard Hickman[1,¶], David Rand[1], David L. Wild[1], Sascha Ott[1], Vicky Buchanan-Wollaston[1,2], Nick Smirnoff[4], Katherine Denby[1,2], Jim Beynon[1,2] and Murray Grant[3]

[1]Warwick Systems Biology Centre, University of Warwick, CV4 7AL, UK
[2]School of Life Sciences, University of Warwick, CV4 7AL, UK
[3]Biosciences, College of Life and Environmental Sciences, University of Exeter, EX4 4QD, UK
[4]These authors contributed equally to this work

Corresponding author: Murray Grant, m.r.grant@exeter.ac.uk

Current Address:
φ The Roslin Institute, Edinburgh, U.K.
α Science and Advice for Scottish Agriculture, Edinburgh, EH12 9FJ
§ Applied Statistics, Rothamsted Research, AL5 2JQ
¶ Department of Plant-Microbe Interactions, Utrecht University, 3584 CH, Utrecht, The Netherlands
# Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Murray Grant (m.r.grant@exeter.ac.uk).

Transcriptional reprogramming is integral to effective plant defence. Pathogen effectors act transcriptionally and post-transcriptionally to suppress defence responses. A major challenge to understanding disease and defence responses is discriminating between transcriptional reprogramming associated with MAMP-triggered immunity (MTI) and those orchestrated by effectors. A high resolution time course of genome-wide expression changes following challenge with *Pseudomonas syringae* pv. *tomato* DC3000 and the non-pathogenic mutant strain DC3000*hrpA*- allowed us to establish causal links between the activities of pathogen effectors and suppression of MTI, and infer with high confidence, a range of processes specifically targeted by effectors. Analysis of this information-rich dataset with a range of computational tools provided insights into the earliest transcriptional events triggered by effector delivery, regulatory mechanisms recruited, and biological processes targeted. We show that the majority of genes contributing to disease or defence are induced within 6h post infection, significantly before pathogen multiplication; suppression of chloroplast associated genes is a rapid MAMP-triggered defence response; suppression of genes involved in chromatin assembly and induction of ubiquitin related genes coincide with increases in pathogen induced ABA; specific combinations of promoter motifs are engaged in fine-tuning the MTI response and active transcriptional suppression at specific promoter configurations by *Pseudomonas*.

3.1 Introduction

Currently the primary methods of disease control against crop pathogens are agrochemical sprays or the deployment of classical plant disease resistance (R) genes using marker-assisted breeding. However, pathogens rapidly overcome most R genes in the field and regulatory changes and a lack of new chemistries have led to a shortage of effective agrochemicals. Therefore, innovative methods need to be developed to provide alternative strategies for crop health (Dangl et al., 2013). One possibility is to re-engineer existing plant defence networks (Grant et al., 2013). A pre-requisite to such an approach is a detailed knowledge of core transcriptional networks recruited during defence and the molecular strategy pathogens deploy to overcome plant innate immunity. A first step towards attaining this knowledge is ensuring a fundamental understanding of how plant-pathogen interactions are propagated at the transcriptional level.

Plants have evolved a robust innate immune system that provides broad-spectrum protection against a variety of pathogens with wide-ranging lifestyles (Jones and Dangl, 2006). Effective plant immunity requires the efficient perception of potentially pathogenic microbial associated molecular patterns (MAMPs) by a range of host encoded extracellular pattern recognition receptors (PRRs) (Belkhadir et al., 2014; Böhm et al., 2014; Macho and Zipfel, 2014; Zipfel, 2014). These stimuli are translated into the rapid transcriptional activation of a network of MTI responses (Buscaill and Rivas, 2014). This germ-line encoded MAMP Triggered Immunity (MTI) provides robust defence against a diverse variety of pathogens and offers biotechnological potential to improve resistance in elite crop varieties. Over the past decade, significant progress has been made in identifying MAMP receptors and their cognate ligands from a range of phytopathogens, including the archetypal PRRs, the leucine rich repeat containing receptor kinase, Flagellin Sensing 2 (FLS2) that recognises bacterial flagellin (Gømez-Gømez and Boller, 2000) and the LYSM domain containing PRR which recognises fungal cell wall chitin (Wan et al., 2008). Combined biochemical and genetic studies are beginning to uncover additional components of these PRR complexes and provide insight into how these receptors are activated and recycled (Göhre et al., 2008; Greeff et al., 2012; Xin and He, 2013; Macho and Zipfel, 2014; Zipfel, 2014). However, a detailed, highly-resolved temporal analysis of the transcriptional responses underpinning MTI and a mechanistic understanding of how these activated networks confer resistance to non-adapted pathogens remains lacking.

Superimposed on PRR activation of MTI is the capacity of pathogens to

produce effectors, comprising both small molecules and proteins. Effectors are usually delivered into the host cell where they target one or more susceptibility factors to attenuate both MTI and effector triggered immunity (ETI) and reconfigure host metabolism to provide pathogen nutrition (Cui et al., 2010; Feng and Zhou, 2012; Cui et al., 2015; Macho and Zipfel, 2015). Effectors act at a number of levels in the MTI signaling cascade to attenuate this defence response (leading to effector-triggered susceptibility (ETS)) including at the PRR interface, or targeting the Golgi, chloroplast, mitochondria or the nucleus (Macho and Zipfel, 2015). Underpinning successful effector driven disease development is a network of complex transcriptional reprogramming events. These early responses collectively overcome MTI and promote pathogen growth. Central to ETS is the pathogen driven modulation of the host hormonal balance, the extent and direction of which appears to be linked to the pathogen's life-style (Robert-Seilaniantz et al., 2011; Pieterse et al., 2012; Kazan and Lyons, 2014). These hormonal perturbations, manifested at both the level of biosynthesis and signalling, are under strong transcriptional control.

Recent comparative genomic sequencing efforts have revealed both reduced and highly plastic pathogen genomes and facilitated the identification of an expanding catalogue of predicted candidate effector proteins, many of which appear to be pathogen lifestyle-specific (Win et al., 2012). With the exception of some conserved pattern motifs, the effector proteins reveal little about their role in pathogen virulence strategies, reflecting the emerging paradigm that they function co-operatively and redundantly, often targeting multiple host components and pathways to successfully promote disease (Lee et al., 2008; Lindeberg et al., 2012; Xin and He, 2013; Macho and Zipfel, 2014). Successful pathogenesis requires suppression of host defence and nutrient acquisition. How this is achieved largely remains enigmatic, including knowledge of the sequence of events necessary to orchestrate disease progression. What is clear is that genetic studies have revealed the existence of many core plant defence components that confer enhanced susceptibility across a broad spectrum of pathogens suggesting that despite deploying disparate effectors, pathogen virulence strategies converge on conserved regulatory hubs (Glazebrook, 2005; Kazan and Lyons, 2014).

Transcriptional reprogramming underpins plant disease and defence strategies. MTI, ETS or ETI networks are all dependent upon transcriptional activation and regulation (Buscaill and Rivas, 2014). Effectors may act post-translationally to modify components of a signaling network, for example, acetylation, or act directly as transcriptional repressors or activators (Lee et al., 2008; Macho and Zipfel, 2014). Thus, understanding the transcriptional dynamics associated with disease develop-

ment affords the possibility of developing precise approaches to interdict pathogen virulence strategies and re-wire host defence responses (Grant et al., 2013). Such approaches first require the capture and interpretation of expression profiles derived from high-resolution sampling of leaves responding to virulent pathogens. Subsequently, to gain an overview of the transcriptional phases of disease development it is necessary to be able to infer how the expression signature of transcription factors can drive or repress expression of downstream network components.

Recently, significant progress has been made towards developing algorithms to mathematically model high resolution RNA expression datasets. However, analysis of gene expression arising from interactions between two biological organisms is challenging and inherent restrictions include insufficient and/or inappropriate time points to provide robust expression profiles for individual genes and to infer statistically significant changes in expression signatures over the infection time frame. An ideal pathosystem will allow confluent and synchronous infection to prevent excessive dilution of signal with uninfected or asynchronous cellular responses. It would provide data to discriminate between transcriptional networks associated with MTI and ETS, providing currently lacking insight into the role of effectors and/or the host response to effector perturbation of innate immunity to be captured (Kazan and Lyons, 2014).

The *Arabidopsis thaliana* — *Pseudomonas syringae* pv. *tomato* DC3000 (DC3000) interaction is ideally suited to dissecting both MTI and ETS processes at the transcriptional level (Xin and He, 2013). DC3000 is highly virulent on *A. thaliana*. DC3000 directly delivers 28 effector proteins (Cunnac et al., 2009) into the host cell through the type III secretion system (T3SS) as well as small molecules such as the phytotoxin, coronatine (Bender et al., 1998). These virulence factors collectively suppress MTI and access nutrients, therefore enabling bacterial multiplication. A key structural component of the T3SS pilus is the HrpA protein (Roine et al., 1997). DC3000*hrpA*- mutants activate MTI but cannot form a T3SS to deliver the suite of effectors required to suppress MTI. They also produce minimal amounts of coronatine (de Torres-Zabala et al., 2009). Thus DC3000*hrpA*- infection triggers MTI in the host. Detailed comparisons between mock, DC3000*hrpA*- and DC3000 treatments capture gene expression associated with MTI, effector-mediated suppression of MTI and subsequent transcriptional changes associated with metabolic reconfigurations that favor pathogen nutrition e.g. deployment of SWEET transporters (Chen et al., 2010, 2012a).

To date, a small number of studies have captured single or limited time points of foliar infections with *P. syringae* (Thilmony et al., 2006; Truman et al.,

2006). These lack temporal context, do not adequately discriminate MTI responses from effector-mediated transcriptional reprogramming and lack corroborating data to link suppression of defence with pathogen proliferation. A mRNA-seq study of virulent and avirulent challenges did not capture MTI, and the combined three time points and complexity of the data limited the ability to interpret phase transitions and/or to temporally pinpoint and monitor key processes (Howard et al., 2013). In this study we generate and analyse high temporal resolution (13 time points) microarray data reporting MTI induced by DC3000*hrpA*- treatment and ETS caused by virulent DC3000 challenge. Inclusion of mock-treated leaves allowed the capture of both gene expression dynamics of MTI induced by DC3000*hrpA*- and how those expression profiles were modulated by DC3000. We provide a timeline for how the host deploys its defence transcriptome, and how components of this may be modulated by effectors. We define sets of promoter elements that potentially drive these changes and using different mathematical approaches define co-regulatory and network models that predict key regulators of plant defence and potential targets of effector manipulation.

3.2 Results

3.2.1 Transcriptional dynamics of MTI and ETS revealed from a large-scale, highly-resolved time series expression study

To design a high-resolution expression experiment we used previous transcriptomic (de Torres-Zabala et al., 2009), proteomic (Jones et al., 2006a,b) and metabolomics (Ward et al., 2010) studies to inform sampling times. We additionally constructed an FRK1 promoter (Flagellin Induced Receptor Kinase 1; AT2G19190) luciferase fusion to ensure the earliest stages of ETS were captured (see Methods). We used high inoculum (OD_{600} 0.15; $\sim$0.75 x 10^8 colony forming units (cfu) mL^{-1}) syringe infiltration of a single fully expanded leaf per plant. Figure 3.1A illustrates that DC3000 but not DC3000*hrpA*- challenge suppressed FRK1:luciferase reporter expression between 3 hpi and 6 hpi. As effector delivery in this system does not occur until $\sim$90 min post infection (Grant et al., 2000), following an initial 0 hpi sample, we then sampled at 2 hpi, and subsequently 3,4,6,7,8,10,11,12,14,16 and 17.5 hpi. For each treatment, leaf 8 from four 34 day-old Col-4 rosettes was sampled, providing 4 biological replicates per time point, per treatment. In addition, four technical replicates were conducted per biological replicate, including a dye swap, thus each treatment at each time point was captured on 16 microarrays, and the experiment used 312 two-colour arrays in total. RNA was prepared and hybridised

to CATMA spotted microarrays (Allemeersch et al., 2005) using a randomised loop array design (Fig. S1) to maximise the comparative power of the experiment. Data was extracted and normalised as previously reported (Breeze et al., 2011) to generate a single expression value for each gene at each time point in each of the three treatments.

To first provide a global overview of transcriptional dynamics across the time course we used scatter plots to represent commonalities and differences in gene expression between DC3000 (ETS) and DC3000*hrpA*- (MTI) challenges relative to the mock control based upon a pair-wise comparison of all genes at each time point. Significantly differentially expressed genes (DEGs) between treatments were determined with the Bioconductor package LIMMA (Smyth et al., 2005) using the Benjamini-Hochberg false discovery rate correction (FDR) and a p-value cut-off of 0.05 (for numbers of DEGs see Supplemental Data Set 1). Scatter plots are used in Fig. 3.1B to illustrate the dynamics of DEGs between DC3000 and DC3000*hrpA*- challenges relative to mock challenge. In these plots green represents DEGs changing in the same direction in both DC3000 and DC3000*hrpA*- challenges, compared with mock ($MgCl_2$) inoculation. Therefore genes categorized as green represent MAMP response genes similarly differentially expressed by both DC3000 or DC3000*hrpA*- challenge. Red represents genes differentially expressed between DC3000 and mock challenge but not between DC3000*hrpA*- and mock challenge. Thus red represents DEGs positively influenced by effectors (greater induction or suppression) relative to their respective DC3000*hrpA*- signature. Conversely, blue represents MAMP responsive genes whose expression is attenuated by DC3000. In summary, the DEGs observed between treatments is represented as such: red, effector enhanced changes; blue MAMP responses suppressed by effectors; green, persistent MAMP responses. Violet, captures DEGs between all three treatments, and these appear late in the time course. Rare black signals indicate genes differentially expressed between all three treatments with opposite trajectories between DC3000 and DC3000*hrpA*- challenges.

As expected, at 2 hpi the majority of DEGs are MAMP responsive. Green and red dominate. The significant red component suggests that effectors are just beginning to have an impact, because at this time point there are no DEGs between DC3000 and DC3000*hrpA*-. Without further experimentation it is not possible to determine whether these transcriptional changes are part of the pathogen virulence strategy or are the result of a weak initial ETI response. A rapid, but transient, change between 2 hpi and 3 hpi captures the first significant effector-driven transcriptional differences between DC3000 and DC3000*hrpA*- challenges. At 3 hpi the

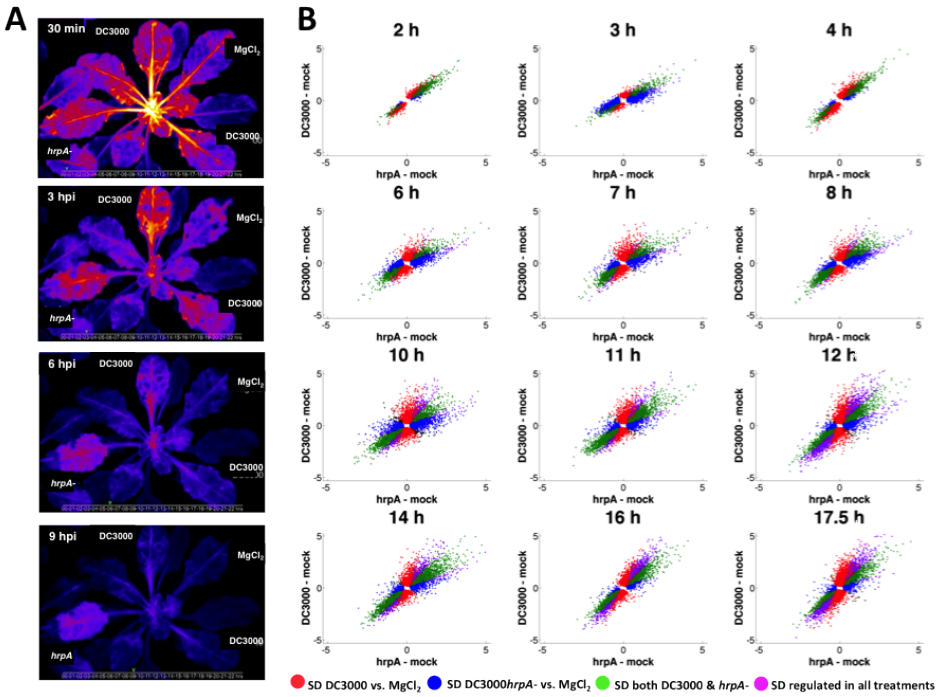

Figure 3.1: Dynamics of differentially expressed genes during basal defence and disease development. **(A)** Infection dynamics reveal suppression of basal defence by DC3000 using a reporter line expressing Flagellin Induced Receptor Kinase 1 (FRK1; AT2G19190) fused to luciferase. Suppression of FRK1 expression is evident between 3-6 hpi following DC3000 challenge, whereas FRK1 expression is sustained in the DC3000*hrpA*- challenged leaf. **(B)** Dynamics of expression in Arabidopsis leaves after challenge with either DC3000 or the DC3000*hrpA*- mutant, representing disease and defence responses respectively. Gene expression is represented graphically at each time-point by a scatter plot. In these plots green represents differentially expressed genes (DEGs) changing in the same direction in both virulent DC3000 and mutant DC3000*hrpA*- challenges, compared with mock ($MgCl_2$) inoculation. Therefore genes categorized as green represent MAMP response genes not modified by effectors. Red represents DEGs between DC3000 and mock challenge but not between DC3000*hrpA*- and mock challenge. Thus red represents genes positively influenced (more strongly induced or suppressed) by effectors relative to their respective DC3000*hrpA*- signature. Conversely, blue represents MAMP responsive genes whose response is attenuated by effectors. In summary, for DEGs in one treatment relative to mock red represents effector enhanced changes relative to DC3000*hrpA*- treatment compared to mock; blue represents MAMP responses modified by effectors; green, persistent MAMP responses. Violet, captures DEGs between all three treatments, and these appear late in the time course. Rare black signals indicate genes differentially expressed between all three treatments with opposite trajectories between DC3000 and DC3000*hrpA*- challenges. Gene expression analysis was performed using the LIMMA (Linear Models for Microarray Data) package in Bioconductor using a p-value cut-off of 0.05 and FDR applied using the Benjamini-Hochberg method.

impact of effectors was striking, despite the amplitude of these responses being relatively small. The strong blue signal indicates DEGs in DC3000*hrpA*- but not DC3000 signifying an initial transcriptional suppression of MTI by effectors resulting in re-alignment of gene profiles in DC3000 infected leaves back to the mock signature. By 4 hpi the profile changes again, with persistent MAMP (green) and DC3000 driven changes (red, corresponding to ETS or weak ETI) dominant and with an obvious increase in amplitude. From 4-6 hpi another major change in MTI and ETS is evident, respectively illustrated by the strong blue and red profiles and appearance of a violet signal. This pattern subsequently consolidates and is characterised by an increasing number of DEGs over time with a diminishing blue signal and the emergence of a strong violet signal. Notably, persistent MAMP responsive expression (green) remains a major component across the time course. This reflects a continued role for activated PRRs in a sustained host defensive response, and highlighting that effectors only modulate a sub-set of MAMP responsive genes.

3.2.2 The majority of transcriptional changes are initiated by 6hpi

To take full advantage of the temporal nature of our datasets and to ensure we captured the breadth of the transcriptional response, DEGs were found using three techniques specifically adapted to time-series data (Fig. S2 and Supplemental Data Set 2): a locally-adapted Gaussian Process (GP) two-sample test modeling time series (GP2S; (Stegle et al., 2010)) which calculates differential expression based on a Bayes factor calculated between two models; one which assumes that the microarray time series in both conditions are samples drawn from an identical shared distribution and an alternative model that describes the time series in both conditions as samples from two independent distributions; Bayesian Analysis of Time Series data (BATS; (Angelini et al., 2008)) which uses the ratios of the expression in the different time series treatments to calculate the Bayes factor indicating whether a gene is differentially expressed, and fits a function to this expression-ratio time series; and MicroArray ANalysis Of VAriance (Wu et al., 2003) which can be applied to the statistical analysis of gene expression data from two-color cDNA microarrays with sophisticated experimental design. These three methods can identify different sets of DEGs and all have advantages and different weaknesses. For example we are confident in the genes identified by GP2S from our previous work (Breeze et al., 2011; Windram et al., 2012). However, due to the process-fitting algorithm the GP2S algorithm is not ideally suited to capture genes that change rapidly in expression as seen in the rapid transitions early in the infection process (Fig. 3.1B). As no method resulted in enriched false positives within the Venn areas in Fig. S2, we

therefore took the union of DEGs predicted by these three methods to allow us to increase the scope of our differential expression analysis. To find the time at which expression of these DEGs first diverges between treatments we applied the Gaussian process gradient tool (Breeze et al., 2011). Figure 3.2 shows the time at which the gradient significantly deviates from zero for the log expression ratios, i.e. the point at which the gradient of the expression ratio significantly increases (up-regulation) or decreases (down-regulation) from zero for mock-subtracted DC3000*hrpA*- expression (Fig. 3.2A; MTI responses), mock-subtracted DC3000 expression (Fig. 3.2B), or effector driven differences in DC3000*hrpA*- subtracted DC3000 expression (Fig. 3.2C). Note, that Fig. 3.2 is labelled with the time at which the gradient starts to change. As many of the transcriptional changes captured in the dataset are rapid MTI related responses, the expression of a significant proportion of genes has already clearly differentiated between treatments by 2 hpi, therefore the gradient appears to be changing from 0 hpi. Comparison of both DC3000 and DC3000*hrpA*- challenges with mock show that the majority of genes exhibit first DE within 6 hpi, with a notable peak evident at 2 hpi, consistent with a strong MTI response. Comparing the time of expression divergence (gradient change) across treatment comparisons, more DEGs were detected during infection with DC3000 (as inferred from Fig. 3.1), but overall both challenges have similar temporal profiles. The impact of effectors increases from 2 hpi (Fig. 3.2C), with a strong peak in DE at 6 hpi, before declining. Between 7-10 hpi, very few new genes exhibit divergent expression in the two treatments for the first time, and virtually none after 10 h. These data illustrate that the majority of the transcriptional response to MTI and ETS is initiated within the first 6 hpi. These signatures are subsequently modified in amplitude and direction or sustained in response to further effector activities during succeeding time points as captured in Figure 3.1.

3.2.3 Early effector activity leads to major transcriptional changes prior to increased bacterial growth

Using the time-series DEGs we next mapped the temporal structure of MTI and effector induced gene expression (Fig. 3.2) on to the respective pathogen growth curves and used gene ontologies (Ashburner et al., 2000) to capture existing knowledge of processes modulated by these contrasting challenges. GO selection was based on minimizing repetitive terms and maximizing informative terms (eg. phytoalexin biosynthesis rather than cellular metabolism). Figure 3.3A illustrates the temporal changes in biological process ontologies (determined using BiNGO (Maere et al., 2005)) enriched in genes DE during MTI (DC3000*hrpA*- vs. mock challenge)

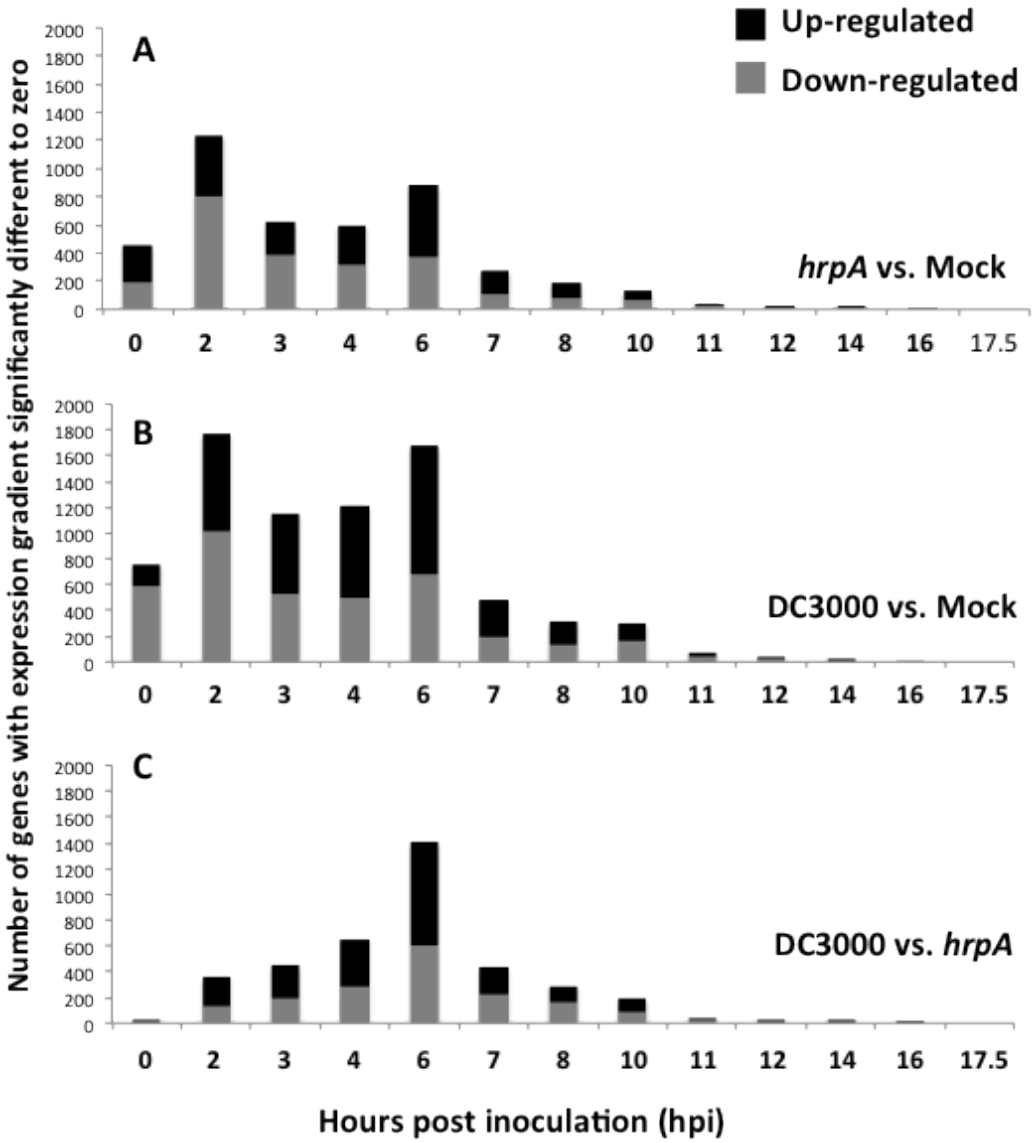

Figure 3.2: Time at which gradients of DEGs begin to significantly differ between treatments. The histograms show the times at which the gradient profile of log expression ratios of DEGs between treatment pairs first diverges from zero as determined by the gradient tool (Breeze et al., 2011). **(A)** mock-subtracted expression during DC3000*hrpA*- infection, **(B)** mock-subtracted expression during DC3000 infection and **(C)** DC3000*hrpA*- subtracted expression during DC3000 infection. Threshold for up/down-regulation is three standard deviations of the gradient being significantly non-zero ($P_{\text{non-zero}} < 0.05$).

mapped onto bar charts depicting DC3000*hrpA*- growth under identical inoculation conditions used for the microarray experiments. Figure 3.3B captures effector modified gene expression (DC3000*hrpA*- vs. virulent DC3000) mapped on to multiplication of the *hrpA* mutant and virulent DC3000, sampling at 0, 4, 6, 7, 8, 9, 10, 12 and 21 hours post infiltration (hpi). Growth curves are annotated with overrepresented gene ontologies of up- (red) or down-regulated (blue) genes separated by the time at which the gradients first diverge between treatments. Bacterial growth curves show that DC3000 does not grow significantly until 8 hpi, whereas DC3000*hrpA*- does not grow during the time course (Fig. 3.3B). Thus the majority of the transcriptional responses to MTI and ETS seen in Figure 3.2 occur prior to multiplication of DC3000. Notably, a reproducible dip in bacterial growth shortly after infiltration can be seen in both DC3000 and DC3000*hrpA*- growth curves and has previously been reported (Mitchell et al., 2015). However, the dip in DC3000 growth appears to be more pronounced, suggesting that the delivery of effectors may initially be detrimental to DC3000 growth, perhaps suggesting a weak but ultimately unsuccessful ETI response within the host. To corroborate the bacterial growth dynamics, confocal images of YFP-expressing DC3000 ($\sim$7.5 X 10^{-7} cfu/ml) within Arabidopsis leaves captured 4, 8 and 22 hpi show the very limited growth of DC3000 at 8 hpi (Fig. 3.3C).

As expected, the early biological process GOs induced by DC3000*hrpA*- challenge represented defence responses. These could be further refined into respiratory burst, phosphorylation, post-translational modification and SA synthesis, consistent with our emerging knowledge of how MAMP receptors respond to their cognate ligands (Kadota et al., 2014; Macho and Zipfel, 2014; Zipfel, 2014). A short time later, at 4 hpi and somewhat counter intuitive given its role in suppression of SA signalling, jasmonic acid (JA) synthesis and response to oxidative stress are over-represented. As we compared syringe-infiltrated, $MgCl_2$ as a control, the JA response that occurs in DC3000*hrpA*- challenged, but not mock-treated leaves, is unlikely to have arisen from the wound response due to the inoculation technique. Interestingly, by 7 hpi, the most dominant ontology is ubiquitin-dependent protein metabolism, a process intimately linked to removal of key regulatory modules (Dudler, 2013). In contrast, a striking enrichment of gene ontologies associated with plastid-targeted genes, in particular photosynthesis related processes, account for the earliest suppressed processes, occurring within 2 hpi. This is dynamic and additional plastid-related genes continue to be suppressed through to 6 hpi. Changes in nuclear-encoded plastid genes have been reported previously in various host-pathogen responses (Bonfig et al., 2006; Truman et al., 2006; Zheng et al., 2012) but these studies have not

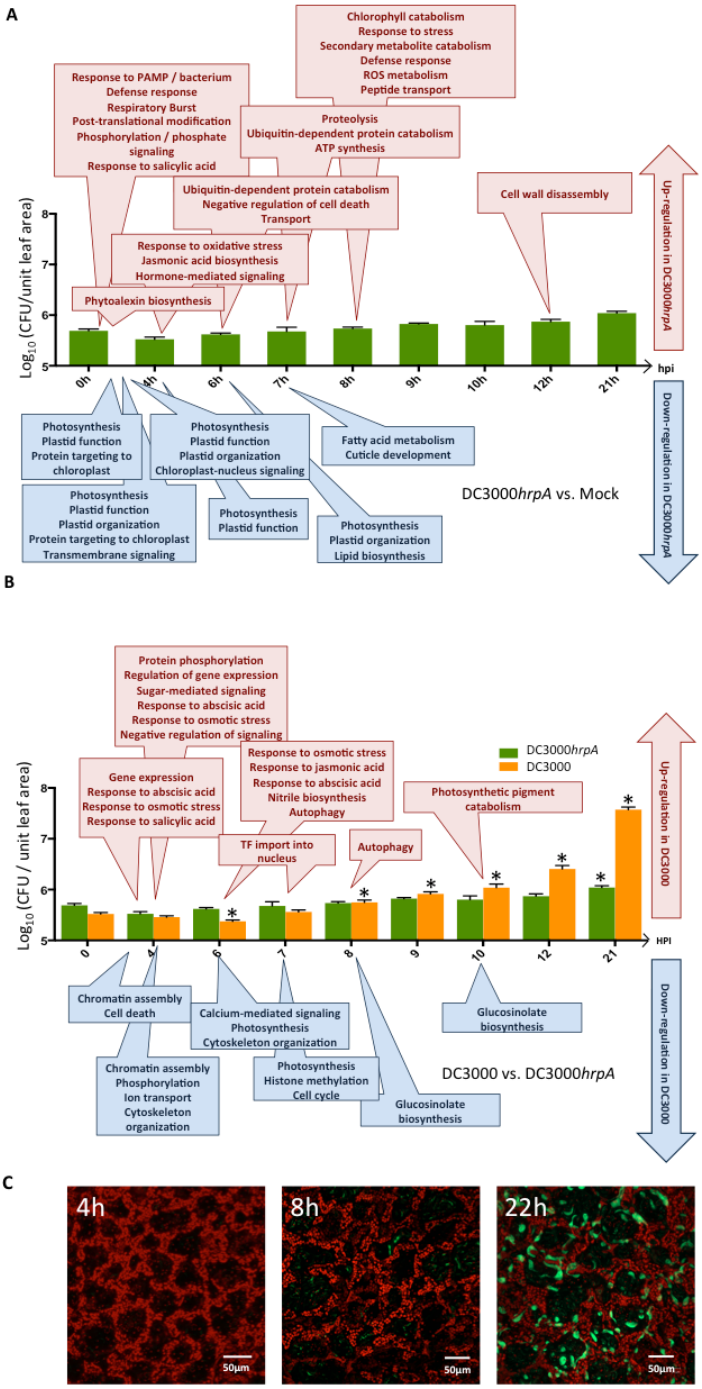

Figure 3.3: Growth curves of DC3000 and DC3000*hrpA*-, with selected GO terms enriched by genes changing expression at indicated time points. Bacterial growth of **(A)** disarmed DC3000*hrpA*- and **(B)** virulent DC3000 following syringe challenge (~0.75 x 10^8 cells/ml). Asterisk represents significance growth differences between treatments as determined by Students t-test ($p < 0.5$, $n = 5$; means $\pm$ SD). Growth curves are annotated with overrepresented GOs of up- (red) or down-regulated (blue) genes separated by the time at which the gradients of DEG profiles begin to deviate (Fig. 3.2). **(A)** ontologies of DEGs between DC3000*hrpA*- to $MgCl_2$ treatments; **(B)** ontologies of DEGs between DC3000 and DC3000*hrpA*- challenges. Gene ontology enrichment was determined using BiNGO (Maere et al., 2005). Growth of YFP-expressing DC3000 within Arabidopsis leaves at 4, 8 and 22 hpi **(C)** corroborates growth curve data.

captured the striking rapidity of this response nor clear attribution to MTI. By 7 hpi, components of fatty acid metabolism and cuticle development (implicating synthesis of waxes) are suppressed. Subsequent to 8 hpi, few other processes are induced, perhaps indicative of the quelling of the energy demanding MTI response in the absence of perceived effector modulation. Selected GO terms enriched in DEGs between DC3000 and mock-infiltrated leaves are shown in Supplemental Figure 3, and the full lists of significant GO terms for all treatment comparisons are given in Supplemental Data Set 3.

Analysis of genes that diverge between DC3000 and DC3000*hrpA*- enables the impact of ETS to be captured and examined in further detail, and in isolation from persistent MTI expression. Notably, ontologies capturing an early induction in host response to ABA between 4 hpi and 6 hpi reinforce the importance of this hormone in suppressing early MAMP responses and promoting susceptibility (de Torres-Zabala et al., 2009). Also of note is the co-ordinated induction of negative regulators of signalling 4 hpi. These are proposed to be a key mechanism in re-configuring host transcriptional responses to virulent pathogens (Kazan, 2006). This is paralleled by over-represented ontologies for transcription factor (TF) import into the nucleus at 6 hpi and autophagy at 7 hpi. Photosynthetic processes, which initially occur as part of MTI in both treatments relative to mock, represents the most dominant effector suppressed ontology. Suppression of photosynthetic processes is maintained throughout the time course in DC3000-infected leaves, whereas recovery of expression can be seen in DC3000*hrpA*- by 17.5 hpi. In addition, chromatin assembly (4 hpi), histone assembly (7 hpi) and glucosinolate biosynthesis (8 hpi) are the most notable ontologies amongst the suppressed genes, suggesting restriction of secondary metabolism and global reconfiguration of the transcriptome architecture. The most striking feature of the effector-induced host transcriptional re-programming is that the majority occurs remarkably rapidly, well before significant bacterial multiplication (Fig. 3.3B).

Although we used 4 single leaf replicates and two technical replicates per treatment per timepoint, this is a single time course. To validate our data we compared the DEGs from this experiment with DEGs from corresponding timepoints and treatments identified by Truman et al. (2006) in a study run under very similar conditions at different locations (Table S1). Strikingly, we saw highly significant concordance between these two studies with a Spearman correlation ranging between 0.76 and 0.90, indicating a remarkable degree of reproducibility between these two experiments and supporting the integrity of these data.

3.2.4 Detailed analysis of gene expression patterns during MTI and ETS

To tease apart the complexity of these responses and identify the groups of genes showing MTI- and effector-responsive expression, we looked at how DEGs (Fig. S2) overlap between pairwise comparisons of the three treatments (Fig. S4 and Supplemental Data Set 4). Each area of the Venn diagram is labeled according to the transcriptional response output it represents. The DEGs in two or more pairwise comparisons (i.e. sections D, E, F and G in Fig. S4) were further subdivided into those up and down-regulated and are illustrated by representative profiles in Figure 3.4. These genes represent four broad response sub-categories. Category D is defined as MTI responsive; these 3880 genes behave similarly in both bacterial treatments relative to mock-inoculated leaves. The ontologies representing transport, ubiquitination, response to bacterium, phosphorylation, response to JA and phytoalexin biosynthesis are induced in Category D whereas supressed genes are enriched in protein import and photosynthesis related components. The large number of genes within this category represent core MTI response components not modulated by effectors (green in the pair-wise comparison in Fig. 3.1) and serves to highlight (i) the scale of the transcriptional reprogramming and (ii) that a successful infection does not require wholesale, but rather selective, suppression of immunity-induced gene expression changes.

Category E contains 525 genes that are responsive to MTI during DC3000*hrpA*- infection, but which are suppressed (by either up or down-regulation), and return to mock expression levels by effector activity during DC3000 infection. Effector suppressed Category E genes are enriched in ontologies capturing phosphorylation, calcium signaling, response to stress, glucosinolate catabolism, cell death and defence responses. Notably, effector-induced ontologies in Category E capture induction of JA and ABA response genes.

A more detailed examination of Category E genes reveal several known regulators of the defence response: the EF-Tu receptor (EFR), TGA3, WRKY53, RBOHD, PEN2 and PEN3. EFR specifically recognizes bacterial elongation factor and activates plant defence responses. It exists in a complex with RBOHD, responsible for the generation of reactive oxygen species during plant defence (Kadota et al., 2014). PEN2 and PEN3 are essential components of cell wall defence linked to metabolism and transport of secondary metabolites (Clay et al., 2009). TGA3 and WRKY53 encode TFs playing critical roles in defence against DC3000 (Kesarwani et al., 2007; Murray et al., 2007). Although it is clearly post-transcriptional events that determine the ultimate activity of these proteins, the suppression of

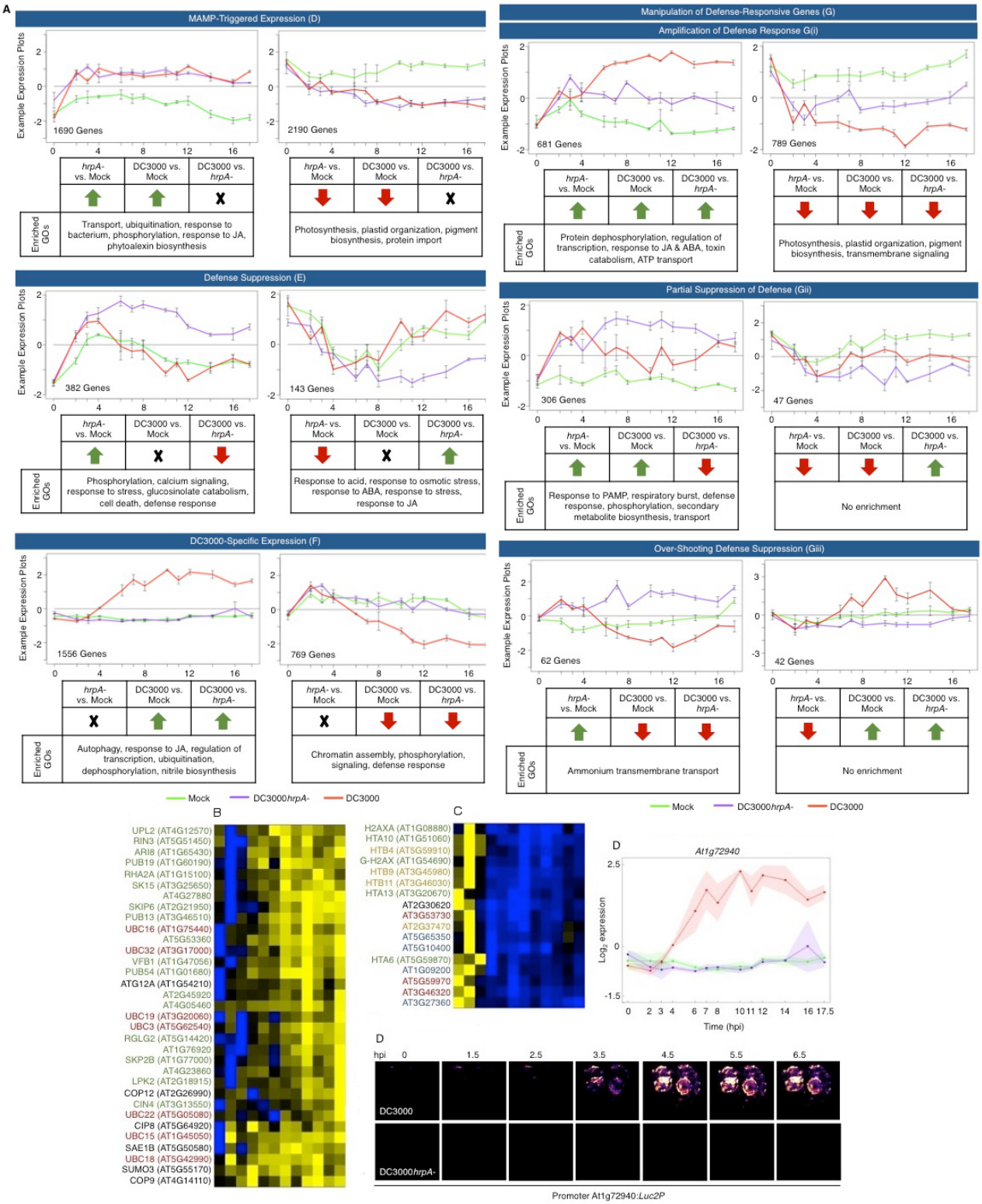

Figure 3.4: Response categories of DEGs capturing different MTI and ETS profiles and their validation

Figure 3.4: Response categories of DEGs capturing different MTI and ETS profiles and their validation. **(A)** Categories derived from the Venn diagram of DEGs (Fig. S4) showing direction of change. Numbers of genes falling into each category with accompanying expression plots (y axis — log relative gene expression, x axis — hpi) for a representative example are shown. Gene ontology enrichments of each sub-category were established using BiNGO (Maere et al., 2005). Heatmaps were generated for chromatin and ubiquitin related genes identified as differentially regulated in Category F (Fig. S4, see Supplementary Data Set 5). Genes were scaled on a per-gene basis and expression represented in blue for genes induced in DC3000*hrpA*- relative to DC3000 and yellow for genes that were significantly higher in DC3000 relative to DC3000*hrpA*-. **(B)** Chromatin associated genes were strongly suppressed in DC3000 challenged leaves. Annotated histone genes are color coded as follows; green = H2A, yellow = H2B, blue = H3, red = H4 and black = H1 linker genes. **(C)** Ubiquitin associated genes differentially regulated in DC3000 challenged leaves. Annotated genes are color coded as follows; red = E2 ligases, green = E3 ligases, black = other related genes (COPs, SUMOs etc). **(D)** Validation of very early effector responsive gene expression using the promoter of the predicted TIR plant disease resistance gene, AT1G72940, fused to a luciferase reporter gene. Transient assay in *Nicotiana benthamiana* showing luciferase activity from 3.5 hpi, following challenge with DC3000 or DC3000*hrpA*-

their transcripts by virulent DC3000 suggests this plays a significant contribution as an effective virulence strategy.

The presence of 6 genes encoding putative Resistance (R) proteins within this group of suppressed defence genes is particularly striking. Five of which encode the TIR class of R protein (TIR-NBS-LRR) and one of the coiled coil type (CC-NBS-LRR). R proteins are a crucial component of ETI and suppression of these 6 putative R proteins by DC3000 effectors suggests that they may play a key role in the plant defence response. Indeed two of these putative R proteins (AT1G12290 and AT4G09420) are able to interact with and potentially 'guard' Arabidopsis proteins that themselves can interact with *P. syringae* effectors (Mukhtar et al., 2011). One hypothesis would therefore be that these R proteins are detecting the activity of pathogen effectors (via the intermediate interacting proteins) and therefore effector mediated transcriptional repression of these R proteins could help to dampen ETI. To our knowledge, recessive *avr* genes, i.e. pathogenic effectors that prevent the recognition of other effectors, have yet to be identified in *P. syringae* pv. *tomato*, although such genes have been found in fungal and other bacterial plant pathogens (Iyer-Pascuzzi and McCouch, 2007).

Another hypothesis emerging from analysis of DC3000-suppressed genes involves protein disulfide isomerase (PDI) activity. Four PDI genes (PDI1, 2, 5 and 6)

are up-regulated during MTI and suppressed by the virulent pathogen, presumably being indirect targets of one or more effectors. These PDI proteins are localized to the ER with at least one, PDI6, also being targeted to chloroplasts (Yuen et al., 2013; Wittenberg et al., 2014). Knockout mutants of PDI6 show reduced photoinhibition due to enhanced repair of photosystems (Garcia et al., 2008; Wittenberg et al., 2014). Hence preventing induction of PDI6 may represent a strategy designed to maintain energy and nutrient supplies.

The signatures of the 2,325 genes in Category F represent those uniquely responsive to DC3000 challenge. Their expression profiles are similar in DC3000*hrpA*- and mock-inoculated leaves, indicating that they are not components of MTI, but that they represent specific effector responsive transcriptional reprogramming. Without additional experimentation, it is not possible to determine whether these are transcribed as part of the pathogen virulence strategy or representative of a host response to effector activities. These genes may be involved in a host secondary defence response, may be responding to bacterial proliferation, or may be susceptibility targets of effectors that facilitate the infection process.

As expected, Category F encompasses a diverse range of genes reflecting the necessarily broad range of processes targeted by effectors, or induced by the host in an attempt to mitigate the extent of effector activities. Ontologies enriched in these effector induced genes include autophagy, response to JA, regulation of transcription, ubiquitination and dephosphorylation, whereas suppressed genes include ontologies associated with chromatin assembly, phosphorylation, signaling and defence response.

The most striking gene class down-regulated in Category F was the family encoding receptor-like-proteins (AtRLP). Arabidopsis has 57 RLPs, 15 of which are down-regulated ∼10 hpi in Category F, consistent with effector mediated attenuation of host defence capability. Two other notable features in this category were a large number of MYB domain encoding genes (30), and 4 out of 6 of the ABI5 binding proteins which regulate ABA signaling (Garcia et al., 2008), including NINJA, the negative regulator of JA signaling (Pauwels et al., 2010).

Looking at biological processes, most notable were the strong suppression of transcripts (Supplementary Data Set 5) associated with chromatin reorganization (Fig. 3.4B) and the induction of components annotated as playing a role in ubiquitination (Fig. 3.4C), particularly as both these processes play broadly complementary roles in controlling transcriptional networks (Dantuma et al., 2006; Zhao et al., 2014).

To validate the expression profiles we made promoter luciferase fusions be-

tween two genes rapidly induced by effectors and tested their response to bacterial challenge by transient assay in *Nicotiana benthamiana*. We chose a classical TIR-NBS-LRR (AT1G72940) and a RAB GTPase homolog C2B GTP-binding transcription factor (AT3G09910). In contrast to the suppression of R genes in Category E and RLP's in Category F above, the strikingly rapid induction of AT1G72940 in response to DC3000 but not DC3000*hrpA*- challenge (Fig. 3.4D) suggests activation of a host defence response to effectors and reinforces the possibility of active engagement of an early ETI response. AT3G09910 showed a similar DC3000 responsive profile in *N. benthamia* transient assays but, consistent with the microarray data, the magnitude of response is significantly less than that driven by the AT1G72940 promoter (Figure S5).

Finally, Category G contains 1927 DEGs that behave differently across all three treatments. Broadly, Category G represents genes involved in the manipulation of defence. These genes can be further classified as follows; 681 and 789 genes with a DC3000 profile that appears to be respectively an amplification or suppression of the DC3000*hrpA*- response compared to mock and most likely to represent a sustained host defence response that serves to restrict pathogen multiplication; 306 and 47 genes respectively that have an intermediate expression profile in that they are suppressed or induced in DC3000-treated leaves, compared to DC3000*hrpA*-, but do not realign to expression levels in mock-infiltrated leaves, indicating a partial quelling of MTI. A further 62 and 42 genes show opposing responses to DC3000*hrpA*- and DC3000 infection, compared to mock. These sub-categories serve to highlight the underlying complexity of the transcriptional response.

3.2.5 Early sustained effector specific DEGs are predicted to modulate perception of external stimuli and chromatin reorganisation

We selected effector specific DEG profiles to probe processes initially targeted by DC3000. We identified 140 potential effector-induced genes and 42 potential effector-repressed genes using the criteria that the gene had a sustained (6-8 hpi) differential expression profile between DC3000 vs mock and between DC3000 vs DC3000*hrpA*-, but not between mock vs DC3000*hrpA*- (Supplemental Data Set 6). These genes are predicted to capture the initial host transcriptional changes driven by effector delivery and not induced during MTI. Unexpectedly, not a single GO term was over-represented in the up-regulated list (corrected $p>0.05$). This is consistent with the hypothesis that effectors target a broad range of host genes and this interval captures early effector action before initial effects propagate through the network to

affect significant numbers of genes in individual GO term categories. By contrast, in the suppressed genes we found highly significant over-representation of the term 'chromatin assembly' (corrected $p<1.4e^{-12}$) with histones H1.2, GAMMA-H2AX, HTA13, HTB2, HTB9, HTB11, H3 and H4 strongly suppressed following effector delivery. Two representative examples are shown in Figure S6. MEME (Bailey et al., 2006) analysis revealed that promoter sequences of five of these eight genes contained a paired Oct-TCA motif in close proximity to the transcriptional start site. This motif, originally identified in tobacco histone genes, has been shown to confer S phase-specific transcriptional activation (Taoka et al., 1999). These data suggest that effectors either directly bind to Oct-TCA and neighbouring motifs or recruit/maintain transcriptional repressors (TR) on the promoters of these histones.

Additional inspection revealed suppression of ten putative defence related receptor like kinase encoding genes (encoding 3 leucine rich repeat receptor kinases, 2 cysteine rich RLKs, 3 RLKs, 2 TIR-domain resistance gene homologues) suggesting rapid ETS of these signaling modules. Our data imply that an early virulence strategy restricts components of chromatin assembly, which would have a global effect on nucleosome packaging, providing enhanced access for transcriptional regulators. Concomitantly, transcripts for putative defence receptors/resistance genes are suppressed to potentially dampen further host defence responses as seen for RLPs in Fig. 3.4A Category F.

3.2.6 Investigation of regulatory elements driving establishment of defence or disease

Conserved DNA sequences upstream of transcriptional start sites represent important regulatory regions of the promoter (Baxter et al., 2012). The combinatorial arrangement of these motifs, the nature of the cognate TF regulation, TF availability, the presence of post-translational modifications and competing TRs collectively determine gene expression. We extended our targeted MEME analysis of early effector specific genes to all expression patterns observed and all DEGs. We employed unsupervised clustering (SplineCluster (Heard et al., 2006)) of expression profiles within each treatment (Supplemental Data Set 7) and then for all clusters analysed motif occurrences (Supplemental Data Set 8). Clustering expression profiles over time lends strength to the hypothesis that genes in a cluster are not only co-expressed but also co-regulated.

A total of 32 clusters across all treatment comparisons showed statistically significant overrepresentation for at least one motif (Fig. 3.5). Variation is observed in terms of both the range of motifs distributed over different comparisons and

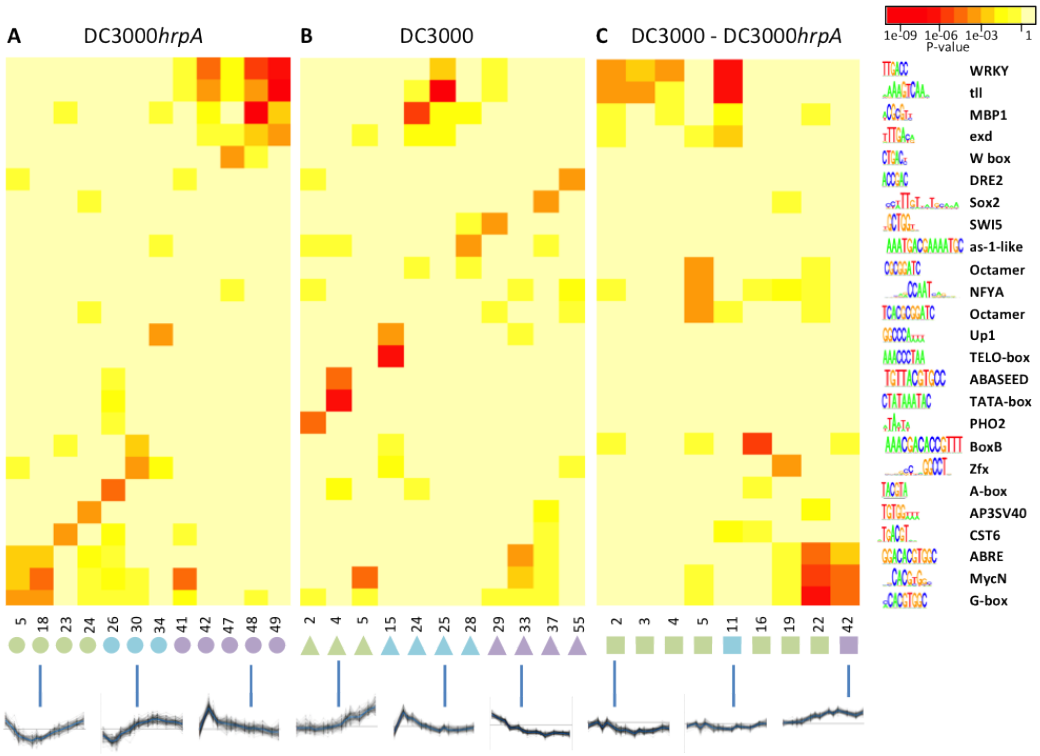

Figure 3.5: Revealing links between TF binding motifs and temporal expression patterns. Over-representation of known TF binding motifs in promoters of gene clusters in three sets of expression clusters. Genes were clustered by **(A)** expression in DC3000*hrpA*-, **(B)** expression in DC3000, and **(C)** expression in DC3000*hrpA*-subtracted from DC3000. Clusters were ordered by profile similarity. Cluster numbers are given on the horizontal axis, colored symbols indicate clusters with similar profiles, and a selected cluster expression profile of each type is plotted below. Names and sequence logo representations of TF binding motifs (where character size indicates nucleotide frequency) are shown on the vertical axis. Colored boxes correspond to p-values. p-values are comparable across rows and columns, i.e. not affected by cluster sizes (see Methods). Rows/columns where at least one cluster-motif pairing shows significant enrichment ($p \leq 1e^{-4}$) are shown (for full results see Supplemental Data Set 8).

different clusters within a comparison. Notably, where the same motif was linked with more than one expression cluster these expression clusters largely had similar temporal profiles, further validating genuine links between motif occurrence and temporal profiles. For example DC3000*hrpA*- clusters 48 and 49 are associated with the same motifs, as are DC3000-DC3000*hrpA*- clusters 2, 3 and 4.

The motifs identified in the DC3000*hrpA*- comparison (Fig. 3.5A) provide an insight into transcriptional regulation of MTI. Genes in clusters 5 and 18 are down-regulated early following DC3000*hrpA*- challenge and contain G-box motifs, whereas clusters 42, 47, 48 and 49 contain W-box motifs and are up-regulated early in infection. This is consistent with the role of WRKYs in the rapid gene activation early in the defence response (Eulgem et al., 2000; Eulgem and Somssich, 2007).

In addition to these motifs, notable plant-specific motifs uniquely over-represented in DC3000*hrpA*- clusters include (i) a specific W-box in cluster 47 involved in elicitor-induced activation of the tobacco chitinase gene which is bound by NtWRKYs 1, 2 and 4 (Yamamoto et al., 2004); (ii) the A-box (TACGTA) which is overrepresented in cluster 26; (iii) the AP3SV40 motif in cluster 24; and (iv) the Zfx box and (v) BoxB in cluster 30.

A number of motifs were identified as strongly overrepresented specifically in DC3000 expression clusters (Fig. 3.5B). In clusters showing suppression these include (i) the DRE2 (drought-responsive element (Busk and Pages, 1998) core element, first described in the promoter of ABA-inducible Rab17, (ii) SWI5 (Badis et al., 2008) and (iii) an as-1-like motif found in the cucumber hydroxypyruvate reductase A promoter, required for cytokinin responsiveness (Jin et al., 1998). Cluster 15, characterized by rapid early gene induction had the Telo-box (Axelos et al., 1989) overrepresented, a motif found in the 5' region of numerous genes encoding components of the translational apparatus. Gene clusters with a late induction profile were characterized by over-representation of the ABASEED motif involved in ABA regulation and seed expression (Busk and Pages, 1998), a specific variant of the TATA-box (cluster 4); or the PHO2 box motif (cluster 2).

The DC3000-DC3000*hrpA*- comparison (Fig. 3.5C) highlights clusters that have combinatorially overrepresented motifs; WRKY, tll, exd motifs in cluster 11 and NFY and two histone OCTAMER motifs in cluster 5 revealing a set of motifs that appear to recruit TFs that are central to ETS. Thus both our targeted analysis of effector suppressed genes and untargeted global motif analysis suggest effectors specifically target a sub-set of genes with OCT motifs. NFY motifs are marginally over-represented in multiple clusters and it is notable that NFY TFs have been shown to recruit histone deacetylase to NFY containing motifs to inhibit promoter

activity (Peng and Jahroudi, 2003) and NFY TFs are largely accepted as providing a fundamental link between chromatin and transcription (Dolfini et al., 2012).

DC3000*hrpA*- clusters 42, 47, 48 and 49 that show a very rapid (2-3hpi) MTI response are enriched in WRKY, tll, MBP1 and exd motifs (Fig. 3.5A). Notably, DC3000-DC3000*hrpA*- clusters 2, 3, 4 and 11 are also enriched with these motifs; they are early-induced in both DC3000 and DC3000hrpA- infections, but in DC3000 infection the expression of genes with these defence-related promoter motifs is later suppressed (Fig. 3.5C). By contrast, ABRE, MycN and G-box motifs are highly over-represented in DC3000-DC3000*hrpA*- clusters 22 and 42 (Fig. 3.5C), comprising genes that are induced by effectors. Strikingly, these motifs are significantly over-represented in genes that are suppressed early in the defence response to DC3000*hrpA*- challenge implying ETS deploys active transcriptional suppression at specific promoter configurations. In summary, we demonstrate a degree of motif specificity and co-operativity in the complex transcriptional regulatory networks recruited during MTI and the modulation of innate immunity by effectors.

3.2.7 Multiple time series co-expression analysis predicts specific regulation of pathogen-responsive genes

As explained above, co-expression of genes has regularly been used as an indicator of co-regulation, and co-expression throughout a high-resolution time series as described here should enhance the likelihood of identifying such co-regulated genes. We extended this analysis by using Wigwams (Polanski et al., 2014) to identify genes co-expressed across at least two of the three time series treatments. In contrast to SplineCluster, Wigwams does not partition the genes into clusters but identifies modules of genes showing statistically significant co-expression (i.e. not all genes in the analysis will end up in a module). The statistical test in Wigwams discriminates between co-expression stemming from the abundance of a particular expression profile and co-expression indicative of co-regulatory mechanisms acting in multiple time series, which is of higher relevance to understanding the underlying gene expression dynamics of the host response. In total, 309 modules (containing 6685 unique genes) were identified that showed statistically significant co-expression in at least two time series (Supplemental Data Set 9). These modules contain genes that may not exhibit expression profile similarities between treatments, but they have similar expression profiles to each other within two or more treatments. For example, the module represented in Figure 3.6A includes genes that all share a pattern of sustained up-regulation in DC3000 and transient up-regulation in the other two time series which, by the Wigwams test, is not a mere consequence of the abundance of such trajec-

tories in the individual time series. In an attempt to identify regulatory events positively or negatively modulating their transcription, modules were evaluated for enrichment of known TF binding motifs in the genes' upstream sequences. A total of 31 modules were enriched for known TF binding motifs targeted by a range of different TF families (Supplemental Data Set 10a). We determined the genes within these modules that contained the enriched TF binding motif(s) in their promoter sequences, hence the most likely co-regulated group, and two examples are shown in Figure 3.6A and B and Supplemental Data Set 10b.

The 17 genes in Figure 3.6A show co-expression across DC3000*hrpA*- and DC3000 infection and all contain a MYB TF binding motif in their promoters suggesting their expression could be controlled by a MYB TF. Intriguingly this group of genes includes three protein phosphorylation enzymes (two kinases and a phosphatase) with known positive effects on defence or defence gene expression; PKS5, WIN2 and CRK45. PKS5 phosphorylates NPR1 and expression of two NPR1-target genes, WRKY38 and WRKY62, is reduced in *pks5* mutants (Xie et al., 2010). Overexpression of the phosphatase WIN2 enhances resistance against DC3000 and WIN2 has been shown to interact with HopW1-1, an effector present in DC3000 (Lee et al., 2008). Similarly, overexpression of the cysteine-rich receptor-like kinase CRK45 enhances expression of defence genes and resistance against DC3000, and mutants lacking CRK45 show increased susceptibility to the pathogen (Zhang et al., 2013). These genes appear to be part of an ultimately unsuccessful effector-triggered immune response driven by a MYB TF. A second group of potentially co-regulated genes is shown in Figure 3.6B. These 54 genes contain a WRKY TF binding site in their promoters and exhibit co-expression across both pathogen infection time series and the mock-inoculated time series. The group features several genes with a role in defence against *P. syringae*. PUB22, PUB23 and WRKY11 all encode negative regulators of *P. syringae* disease resistance (Journot-Catalino et al., 2006; Trujillo et al., 2008) with FRK1 and WRKY29 key genes induced in response to infection ((Asai et al., 2002) Fig. 3.1A). Furthermore, WRKY29 induces expression of itself and FRK1 (Asai et al., 2002), validating their membership of the same Wigwams module, and suggesting that WRKY29 may regulate the other genes in this group.

Conserved non-coding sequences usually encompass multiple TF binding site suggesting that combinatorial activity of TFs is often required for gene induction or repression (Baxter et al., 2012). We searched for Wigwams modules that were enriched for motifs bound by two or more different TF families and hence, potentially subject to combinatorial regulation. Two examples of such sets of genes are shown in Figure 3.6C and D and Supplemental Data Set 10b. Figure 3.6C fea-

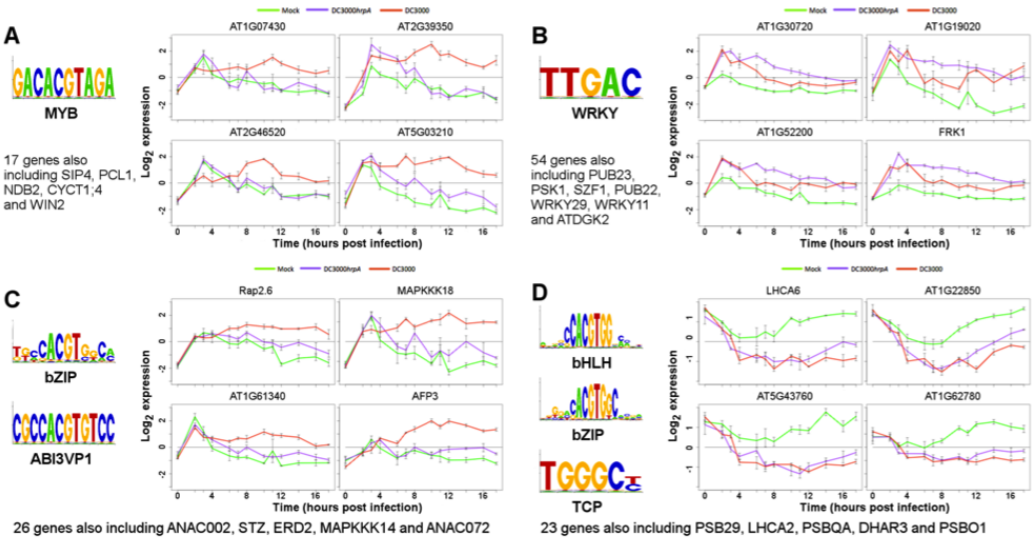

Figure 3.6: Genes containing the same transcription factor binding site(s) in their upstream promoter sequences are co-expressed across multiple conditions. Wigwams modules containing genes showing statistically significant co-expression across at least two of the three conditions were tested for enrichment of TF binding motifs in gene promoter sequences. Genes containing enriched motifs in their promoters were identified. **(A)** Genes co-expressed during DC3000*hrpA*- and DC3000 infection and containing a MYB TF binding motif (PLACE: S-000355) in their upstream 500 bp sequences **(B)** Genes co-expressed during DC3000*hrpA*- and DC3000 infection and containing a WRKY TF binding motif (PLACE: S-000390) in their promoters. **(C)** and **(D)** show examples of genes from Wigwams modules enriched for motifs bound by different families of TFs suggesting combinatorial TF activity regulates expression of these genes. **(C)** Genes co-expressed during DC3000*hrpA*- and DC3000 infection and containing a bZIP binding motif (M00441) and an ABI3VP1 binding motif (S-000145) in their promoters. **(D)** Genes co-expressed during DC3000*hrpA*- and DC3000 infection containing bHLH (M00435), bZIP (M00442) and TCP (S-000474) binding motifs in their upstream Promoter sequences. In all cases, the mean expression profile of representative genes is shown (green, mock; purple, DC3000*hrpA*-; red, DC3000) with error bars indicating standard deviation. The binding motifs, relevant TF family, and names of key genes are provided.

tures 26 genes co-expressed in the mock and DC3000 time series that contain motifs bound by bZIP and ABI3/VP1 TFs in their promoters. This group includes several known ABA inducible genes such as Rap2.6 (Zhu et al., 2010), AFP3 (Garcia et al., 2008), STZ (Sakamoto et al., 2004), ANAC072 (Tran et al., 2004) and MAPKKK18 (Menges et al., 2008). Furthermore, ATAF1 and ANAC072 are components of the ABA signaling pathway and affect the response of the plant to ABA (Fujita et al., 2004; Jensen et al., 2008). The Arabidopsis ABI3 TF and its monocot orthologue VP1 are known to play essential roles in ABA-dependent responses. Although ABI3/VP1 is reported to be seed-specific, there are 14 members of this TF family in Arabidopsis (Riechmann et al., 2000) and other members may play a similar role in other tissues. The hypothesis that co-expression of these genes depends on binding of two TF families is strengthened by the identification of a bZIP protein in rice that interacts with VP1 and mediates ABA-dependent gene expression (Hobo et al., 1999). The final example (Fig. 3.6D) is a group of 23 genes co-expressed in the DC3000 and DC3000*hrpA*- infections and containing bZIP, bHLH and TCP motifs in their promoter sequences. This group contains a number of photosynthetic genes such as PSB29 (Keren et al., 2005), LHCA6 (Ifuku et al., 2005), PSB27 (Chen et al., 2006), PNSL2 (Ifuku et al., 2011), PSBQA and PSBO1 (Murakami et al., 2005). Photosynthetic genes are known to be down-regulated in response to many environmental stress conditions including *P. syringae* challenge ((Bonfig et al., 2006; Truman et al., 2006; de Torres-Zabala et al., in press), Fig. 3.3) and this finding suggests the down-regulation is coordinated by TFs from different families with different DNA binding domains. Select phytochrome interacting PIFs (bHLH TFs) and TCPs have recently been shown to have a role in promoting *P. syringae* multiplication (Weßling et al., 2014).

3.2.8 Modelling the transcriptional network topology during disease and defence

With the exception of the module in Figure 3.6B, the Wigwams analysis above predicts the family of TF regulating a module but does not elucidate specific TF-target gene interactions. Our high-resolution time series experiment was specifically designed to apply network inference approaches, allowing predictions of putative causal regulatory interactions between genes. As TFs determine the topology of transcriptional networks, we inferred regulatory interactions between TFs differentially expressed early during DC3000 and/or DC3000*hrpA*- infection using the Causal Structure Identification algorithm (CSI; (Penfold and Wild, 2011)). We modeled TFs differentially expressed at or before 8 hpi to capture the key early

events associated with establishment of MTI and activation of ETS. These time points also ensured that expression data are not confounded by increased bacterial multiplication in the compatible interaction. We anticipate that the network topology (i.e. specific TF-promoter cis element interaction) is the same during challenge with either DC3000 or DC3000*hrpA*-, but that the flow of information through the network will differ. Specific TFs, their subcellular location, post-translational modifications, and expression levels will determine information flow during the different bacterial challenges. The CSI algorithm is capable of identifying connections between genes that are only 'active' in a subset of the available datasets, providing the expression profiles in the other datasets do not contradict them. Hence network inference was performed using the expression data for the mock, DC3000 and DC3000*hrpA*- infections simultaneously.

The resulting network model features 609 interactions between 433 nodes thus representing a relatively sparse interaction interface (Supplemental Data Set 11). These more linear associations may be indicative of a strategy to rapidly activate TFs driving specific sub-networks rather than using an extended transcriptional cascade. Such a topology would add robustness and make pathogen intervention approaches more challenging. A similar sparse relatively independent network topology was inferred in a recent 'global' model of effector-triggered immunity derived from a range of disparate datasets (Dong et al., 2015). The network model is shown in Figure 3.7 with node color illustrating the differential flow of information through the network in response to DC3000 or DC3000*hrpA*- infection, and the impact of effectors. Expression at a single time point is shown in Figure 3.7 and a movie capturing the dynamic changes in expression of the whole network can be found in Supplemental Data Set 11. To facilitate visualization and interpretation, genes are classified into 8 groups, referenced 1-8 clockwise, according to the difference in their expression between DC3000 and DC3000*hrpA*- infection and containing respectively 3, 23, 59, 87, 53, 92, 81 and 35 nodes. This arrangement is designed to highlight the complexity of network topologies driving the multiple signaling outputs and the inherent host MTI and ETS responses. In Figure 3.7A, the network nodes are colored according to the difference in expression level between DC3000*hrpA*- and the mock with red indicating a TF is up-regulated after DC3000*hrpA*- challenge compared to mock, and green indicating down-regulation. Figure 3.7B shows the same network with the nodes colored according to the difference in expression level after DC3000 infection compared to mock inoculation, and Figure 3.7C shows the network with nodes colored according to the difference in expression between DC3000 and DC3000*hrpA*- infection; nodes with higher expression after DC3000 infection

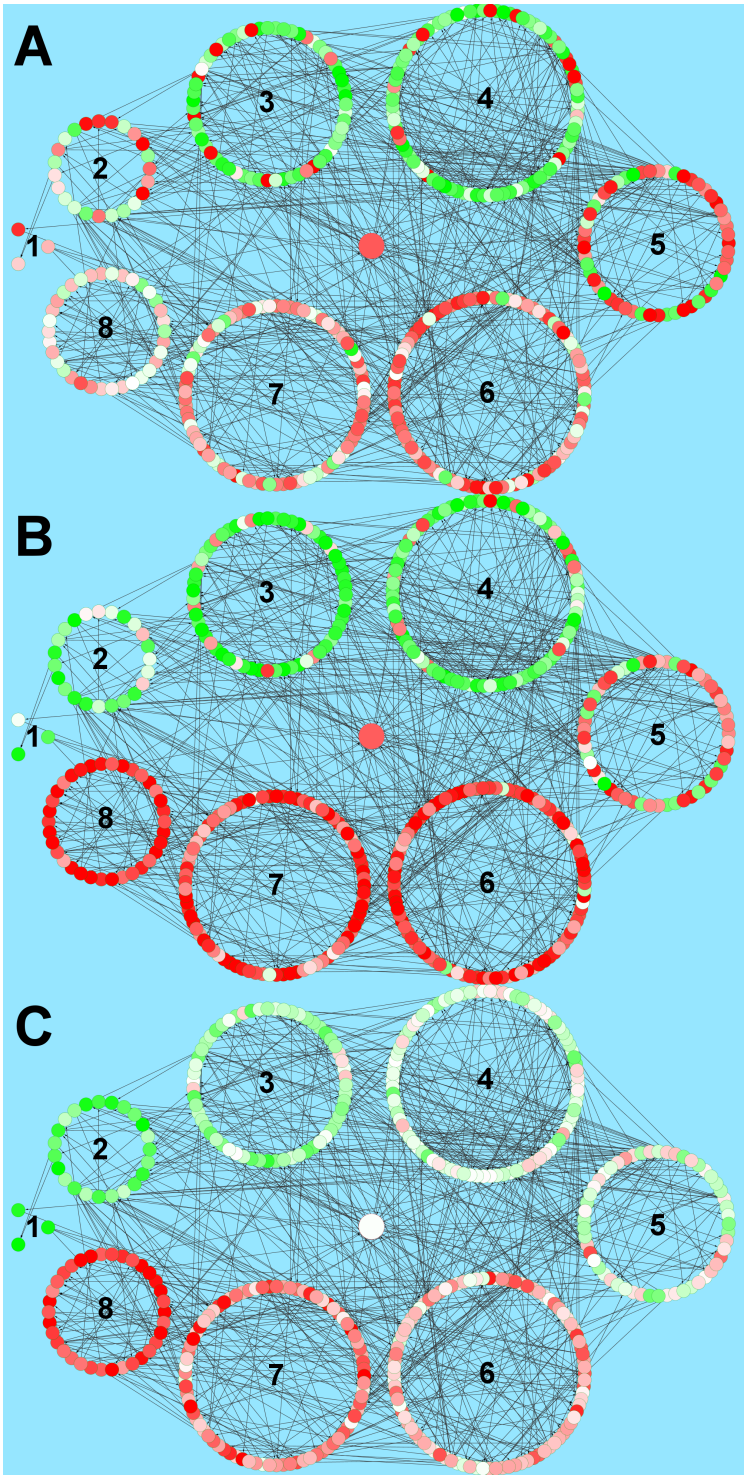

Figure 3.7: The inferred transcription factor network model, jointly obtained for mock, DC3000*hrpA*- and DC3000. The model is limited to genes deemed differentially expressed in at least two of the three pairwise differential expression comparisons, and additionally showing an early response. The visualisation shows the expression levels of the genes in the network at 8 hours post infection in a comparison between **(A)** DC3000*hrpA*- and mock, **(B)** DC3000 and mock, and **(C)** DC3000 and DC3000*hrpA*-. Higher expression levels in the former condition always correspond to red coloured nodes, whilst green nodes represent higher expression in the latter condition. For ease of viewing, the genes in the network were grouped based on their expression trend in the DC3000 and DC3000*hrpA*- delta profile, as shown in **(C)**.

are colored red and those with higher expression following DC3000*hrpA*- challenge are colored green.

The network visualisations in Figure 3.7A and 3.7B demonstrate that the majority of the network TFs are expressed in a similar manner (i.e. up- or down-regulated) in response to DC3000 or DC3000*hrpA*- infection, although expression is also modulated by treatment (Fig 3.7C), as inferred above (Fig. 3.1-3.4).

Figure 3.7C (and most notably the movie in Supplemental Data Set 12) captures network flux in response to effector delivery and groups 1, 2 and 8 highlight markedly contrasting gene expression responses to the presence of effectors. White nodes are similarly induced by DC3000 or DC3000*hrpA*- challenge reflecting that a significant component of the MTI network is not actually perturbed by effectors. This is particularly pronounced in network groups 4, 5 and 6, whereas groups 3 and 7 comprise a mixture of MTI responsive and effector modulated components.

Each network group has numerous interesting components and we will highlight a few of these. Group 3 (59 members) is characterised by 6 auxin response factor/IAA TFs, 9 homeobox domain containing TFs and 6 WRKY TFs. Notably all WRKY TFs (WRKYs 9, 11, 17, 22, 38, 48 and 54) are strongly induced by DC3000*hrpA*- but not DC3000, suggesting these are key defence targets transcriptionally attenuated by effectors. The majority of Group 4 components (87 TFs) were associated with MTI (Fig. 3.7C). Interestingly there was a strong representation of genes related to floral and leaf development including 5 CONSTANS like genes, a regulator of CONSTANS, CAULIFLOWER, AINTEGUMENTA, ETTIN (ARF3), REDUCED VERNALIZATION, AGAMOUS-LIKE87, BELLRINGER, LATE ELONGATED HYPOCOTYL, MERISTEM LAYER 1 and ASYMMETRIC LEAVES 1 as well as chloroplast localized RNA POLYMERASE SIGMA FACTORS 1, 2, A and F. Down-regulation of these TFs associated with developmental processes appears to be a core MTI response, unaltered by effectors, underlining the inter-relationship of transcriptional regulation in development and innate immunity.

Group 2 consists of genes strongly suppressed by effectors. Of the 23 nodes, more than 25% (6) represent MYB domain containing TFs (reinforcing the strong MYB domain representation seen in Category F). These comprise a number of negative regulators of transcription including RAV1 and RAV2, and TCP20 (a negative regulator of senescence and cell size (Lopez et al., 2015)). Thus a strategy for rapid transcriptional activation of pathways by effectors may be to suppress negative regulators of susceptibility targets.

Crucially we can use this network to predict the specific action of pathogen effectors during DC3000 infection. For example, the TFs in group 8 are strongly

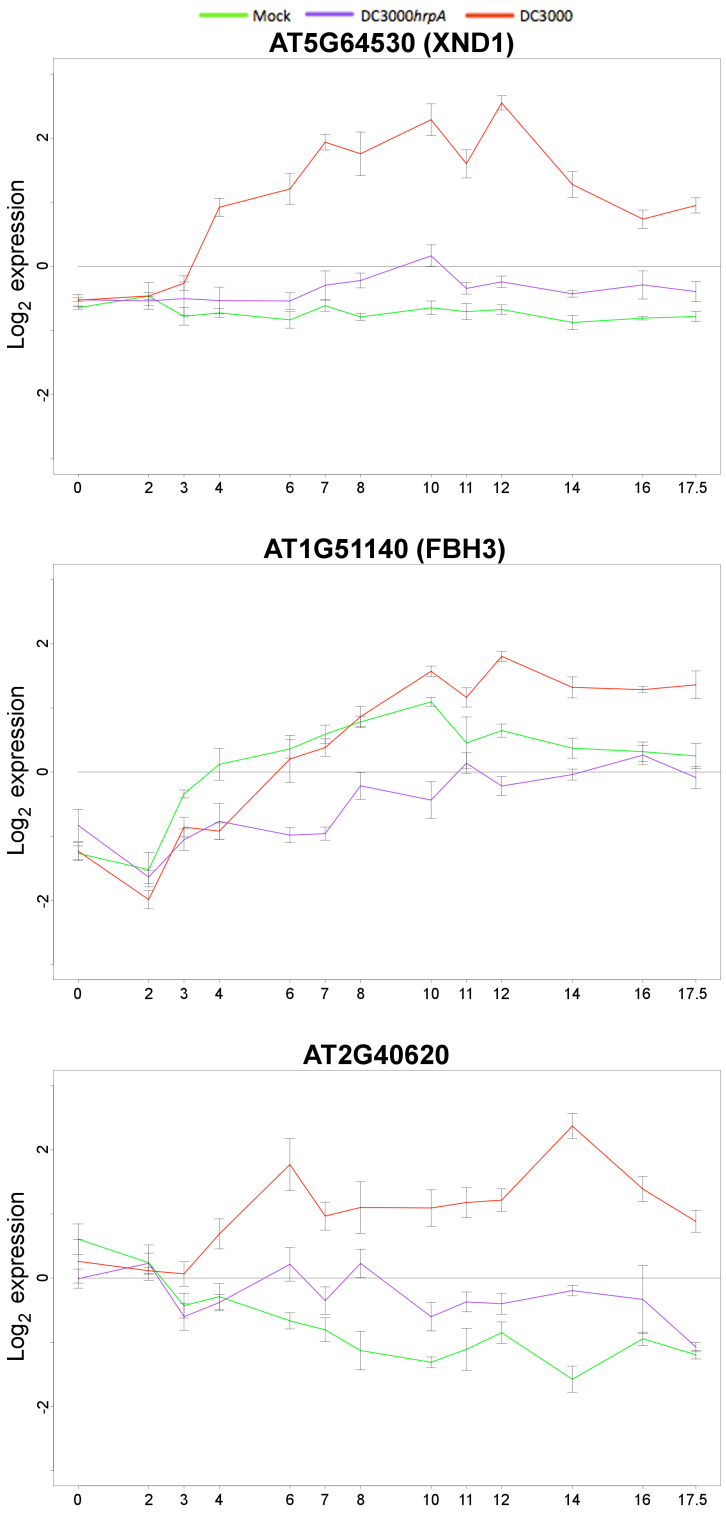

Figure 3.8: The expression profiles of three genes present in the inferred transcription factor network model. Genes were identified as having a high number of downstream targets among genes up-regulated in DC3000 response whilst being down-regulated in DC3000*hrpA*- response. The identified genes are XND1 (AT5G64530), FBH3 (AT1G51140) and AT2G40620. XND1 is thought to negatively regulate cell death, FBH3 contributes to early flowering and AT2G40620's function is currently unknown.

induced following infection with DC3000, whereas they show minimal change compared to mock after DC3000*hrpA*- inoculation suggesting that pathogen effector proteins drive their expression. We analysed the network to identify TFs that were predicted to regulate several of these effector-activated TFs. We hypothesize that effectors cause misregulation of these upstream TFs, which subsequently up-regulate their downstream TF target genes. Three TFs predicted to regulate a number of the TFs in group 8 are XND1 (AT5G64530), FBH3 (AT1G51140) and AT2G40620, all of which show a remarkably rapid change in expression specifically in response to DC3000 infection around 3-4 hpi (Fig. 3.8). This coincides with the timing of the first effector-driven changes in gene expression and is consistent with these three TFs playing a key role in mediating effector influence within the plant.

XND1 is a member of the NAC (NAM (no apical meristem), ATAF, CUC (cup-shaped cotyledon)) TF family and thought to negatively regulate programmed cell death in xylem cells (Zhao et al., 2008). This would be consistent with its predicted effector-mediated up-regulation as a hemi-biotrophic pathogen would benefit from suppression of cell death. FBH3 encodes a basic helix-loop-helix TF that can activate expression of the CONSTANS gene and cause early flowering (Ito et al., 2012). Early flowering in response to *P. syringae* infection has been observed (Korves and Bergelson, 2003) but whether this is a developmental response driven by the plant (as a means to escape disease) or driven by the pathogen is not clear. Our findings are consistent with the active manipulation of flowering within the host plant. AT2G40620 encodes a bZIP TF of unknown function. Our network analysis suggests that these three TFs could represent key hubs playing a major role in the promotion of susceptibility by virulent *P. syringae* (as mediators of effector-driven transcriptional change) and would be candidates for network disruption studies. The mechanism by which these three TFs are up-regulated in response to effectors is not clear, yet their position in the network (upstream nodes) suggests that effector(s) either target their promoters directly or more likely interact post-transcriptionally with other TFs to regulate expression of these hubs. TFs whose activity is determined post-transcriptionally would not be detected in our transcriptional network model. Linking our regulatory network to specific effectors via effector-host protein interaction will be the focus of future research.

3.3 Discussion

3.3.1 The chronology of effector-mediated transcriptional reprogramming

Re-programming of key components of host transcription, facilitating suppression of plant immune responses and acquisition of sufficient carbon and nitrogen resources for bacterial multiplication, underpins successful infection by pathogens. Here we describe a detailed comparative analysis of transcriptional responses in *P. syringae*-infected Arabidopsis leaves.

The primary objectives of this study were to move knowledge beyond single and dual time point studies and (i) capture and contrast the transcriptional dynamics associated with MTI and how these were modified during ETS, (ii) reveal new insights into bacterial virulence mechanisms from the complement of genes targeted by pathogen effector activities during a susceptible interaction and (iii) to provide high quality datasets that can be used by researchers to explore specific transcriptional sub-networks associated with MTI and ETS in detail. To achieve these goals we sampled 4 single leaf biological replicates at 13 time points. Including both technical replicates and dye swaps we probed 16 arrays for each treatment at any specific time point resulting in a very detailed and highly replicated infection dataset comprising 312 two-colour arrays in total. No plant inoculation method is ideal. Dipping or spraying has the advantage of addressing stomatal immunity but disadvantages include the use of high pathogen inoculums and gross differences in bacterial load across the individual leaf sampled. As pathogens enter via the stomata there are spatial differences in pathogen distribution and asynchronous infections, confounding any interpretation of time course data. By contrast, syringe challenge with a defined inoculum concentration ensures (i) infection is as synchronous as possible, avoiding confounding gene expression values, (ii) as much foliar tissue as possible is exposed to the bacterial treatments to maximise response signatures and (iii) leaves of identical developmental stages can be challenged to mitigate confounding developmental impacts. However, the possibility that interactions between wounding from the syringe inoculation and the infection process, or of common responses being masked, remains.

Analysis of this information-rich dataset with a range of computational tools provided insights into the earliest transcriptional events triggered by effector injection, regulatory mechanisms recruited, and biological processes targeted. Central to our study was the ability to relate transcriptional changes to *in planta* bacterial growth. Effector-driven transcriptional modulation was evident as early as 3 hpi,

consistent with *in planta* suppression of luciferase activity seen in leaves of DC3000 challenged FRK1-luciferase lines (Fig. 3.1). Many genes, particularly those responding to DC3000 challenge, showed complex patterns of regulation, with many expression profiles showing very early perturbations from their initial trajectories. Strikingly, despite more than a third of the transcriptome being differentially expressed across our time course, the majority of transcriptional responses (measured by the time at which gradients of expression significantly diverge from mock challenge) were initiated within the first 6 hpi (Fig. 3.2). This is ∼4 hours after effector delivery (Grant et al., 2000) and significantly before measurable increases in bacterial growth at 8 hpi (Fig. 3.3). These early transcriptional changes and complex dynamics during initial effector-mediated transcriptional re-programming have not been captured in previous studies constrained by resolution and sampling strategy.

Within 2 hpi, the impact of effectors was evident in the comparisons between mock and DC3000 or DC3000*hrpA*- challenges (Fig. 3.1), although at this early stage there were no DEGs between DC3000 and DC3000*hrpA*- challenges. By contrast at both 3 hpi and 4 hpi substantial changes in transcriptional dynamics between DC3000 and DC3000*hrpA*- treatments were evident, consistent with early effector activity targeting multiple points of activated PRR signalling pathways (Macho and Zipfel, 2015). By 6 hpi a persistent pattern began to emerge, characterized by an increasing magnitude and number of DEGs between all treatments over time as initial effector targets activated specific transcriptional networks. However, is it important to note that not all effector mediated transcriptional responses are necessarily part of a pathogen virulence strategy. Some may be associated with early, or failed, ETI responses. The early effector responses seen at 2 and 3 hpi with DC3000 in Fig. 3.1, the significantly reduced growth at 6 hpi (Fig. 3.3 and (Mitchell et al., 2015)) and rapid induction of a TIR-NBS-LRR transcript (AT1G72940, Fig. 3.4D) are evidence for an early ETI response.

Approximately 4,000 MAMP responsive genes were identified across the time-course (Fig. 3.4) and represented a major component of the compatible response. These results indicate that while MTI responses are remarkably complex, it is not necessary for effectors to suppress the entire MAMP-responsive network, but rather effectors can selectively suppress specific components or sub-networks to successfully promote disease. This is consistent with the hypothesis that effector target proteins are key components of the host immune response.

The earliest MTI transcriptional response sampled, occurring between 0 and 2 hpi, captured ontologies associated with the respiratory burst, phosphorylation, post-translational modifications and SA synthesis (Fig. 3.3). One of the most

prominent MTI responses was a rapid and sustained suppression of a large proportion of transcripts encoding photosynthetic components, occurring within 2 hpi and being sustained for the first 10 hpi. Thereafter, the majority of these transcripts in DC3000*hrpA*--treated leaves returned to mock levels by 17.5 hpi, whereas in DC3000-treated leaves transcripts largely remained suppressed. These profiles support an increasing belief that chloroplasts are potential integrators of plant immune responses (Stael et al., 2015), and recent experimental evidence in support of this was derived using a subset of the data presented here (de Torres-Zabala et al., in press).

The impact of effectors was early, regulating transcripts encoding a diverse range of proteins. Gene ontologies highlighted ABA biosynthesis as one of the earliest processes induced by effector activity, in agreement with our previous studies (de Torres-Zabala et al., 2007, 2009). In parallel, transcriptional regulators were over-represented amongst early DC3000-induced genes, consistent with an active suppression of MTI transcription responses (Li et al., 2015) and reflected by over-representation of ontologies associated with autophagy (Patel and Dinesh-Kumar, 2008) and TF import into the nucleus coincident with initial bacterial multiplication ∼8hpi.

3.3.2 Biological processes impacted by effectors

Two general patterns of effector-modulation of host transcription were evident. Effectors could modify MAMP-responsive expression through repressing accumulation or preventing suppression of transcripts. We identified 525 MTI responsive genes that in DC3000 challenge were reconfigured by effectors to a mock profile (Fig. 3.4). Amongst these genes were previously well-characterized regulators of defence responses including EFR, TGA3, WRKY53, RBOHD, PEN2 and PEN3. The vast majority of effector-modulated gene profiles were represented by the specific induction or suppression of a diverse variety of 2325 host genes, which were not MAMP responsive (i.e. same profile in mock and DC3000*hrpA*- challenges; Fig. 3.4). Notable was the over-representation of genes involved in transcriptional repression and ubiquitination (Fig. 3.4C), genes encoding MYB domains and the repression of genes encoding plant receptor-like kinases. We hypothesize that the co-ordinated transcriptional reconfiguration of these processes represents a fundamental pathogen virulence strategy. Repression of plant receptor-like kinases to attenuate PRR signaling, or induction of ubiquitin ligase activity encoding genes to facilitate targeted proteolysis are intuitive strategies. However, most striking was the differential regulation of a range of genes involved in chromatin remodeling.

3.3.3 An emerging role for chromatin remodeling early in the susceptible interaction

The earliest sustained transcriptional responses specifically up-regulated by DC3000 but not by DC3000*hrpA*- showed no statistical overrepresentation of any GO functional category term. In stark contrast, early down-regulated genes showed a highly significant enrichment for 'chromatin assembly'. This indicates that at early infection stages, effector activities have not sufficiently propagated through the network to converge onto sufficient numbers of genes in any functional category to yield a statistical signal, whereas down-regulation of chromatin assembly genes was rapid (Fig. 3.4B). Consistent with these findings, motif analysis of promoters of early down-regulated genes identified the octomer-TCA motif as enriched in these genes, and particularly in the subset of chromatin-related gene promoters. The link between this motif and DC3000-specific down-regulation was also reinforced in an unbiased analysis of all temporal expression clusters (Fig. 3.5C).

Covalent modifications or 'marking' of histone tail residues is crucial for regulation of gene transcription within the chromatin context. Histone lysine methyltransferases (HKMTs) and histone lysine demethylases (HKDMs) act collectively to impart specific methylation patterns on particular histone lysine residues (Berr et al., 2011; Zentner and Henikoff, 2013). A large number of histone encoding genes were suppressed by effectors including class 3 and 4 Histones as well as Histone 2A (HTAs 10, 13 and 6) and Histone 2B (HTAs 4, 9 and 11). Notably HTA9 encodes DECREASED DNA METHYLATION 2/cytosine methyltransferase 1 (DDM2/MET1) (Kankel et al., 2003) suggesting that effectors are actively remodelling chromatin early in the infection process. Reprogramming of chromatin remodeling was reinforced by specific analysis of early (6-8 hpi) effector modulated genes, identifying Histones H1.2, GAMMA-H2AX, HTA13, HTB2, HTB9, HTB11, H3 and H4 among the 42 genes suppressed early, with the majority of the others being a mixture of immunity related receptors (Supplemental Data Set 6).

In parallel to repression of numerous histone encoding genes, effectors also induced some chromatin remodelling genes. The induction of HDA15, encoding a histone deacetylase, which negatively regulates chlorophyll biosynthesis and photosynthesis gene expression in etiolated seedlings (Liu et al., 2013) would be likely to contribute to the sustained suppression of photosynthetic genes observed in Figure 3.3. Additionally the SET-domain containing genes KRYPTONITE (SDG33) and ASH1 (SDG26) were induced early in infection. These encode H3K9MTs implicated in cross talk between DNA and histone methylation (Du et al., 2014) and in transcription regulation and flowering time control (Berr et al., 2015) respec-

tively. Thus, the suppression of histone encoding genes and the selective induction of genes involved in covalent modification of histones evidenced from this study is consistent with an early pathogen strategy to render stretches of DNA more accessible to TFs. It is particularly notable that these changes coincide with increased ABA ∼6 hpi (de Torres-Zabala et al., 2007, 2009). Mechanistic links to chromatin remodeling and ABA (Chinnusamy et al., 2008; Ma et al., 2011) or deubiquitination (Sridhar et al., 2007) are emerging. The Arabidopsis homolog of the yeast SWI3 subunit of SWI/SNF (Switch/Sucrose Nonfermenting) chromatin-remodeling complexes (Sarnowski et al., 2002) SWI3B, interacts with the PP2C, HAB1. *swi3b* mutants show reduced sensitivity to ABA, implicating SWI3B as a novel positive regulator of ABA signaling regulated by HAB1 (Saez et al., 2008). By contrast, the Arabidopsis ATPase BRAHMA SWI2/SNF2 complex represses the ABA INSENSITIVE5 (ABI5) bZIP transcription factor which is important in regulating a range of ABA mediated functions including vegetative growth and water stress responses (Han et al., 2012).

3.3.4 Predictions of regulatory relationships underlying MTI and ETS

To gain further insight into how differential MTI and ETS transcriptional signatures evolved we looked for individual and combinatorial TF binding motifs in promoter elements, first using unsupervised clustering to identify DEGs that exhibit similar treatment-specific expression profiles then screening these clusters for overrepresentation of known TF binding motifs. In addition to previously reported WRKY boxes in early-induced genes we also found G-boxes enriched within clusters of suppressed genes and enrichment of specific motifs in gene clusters showing opposing expression profiles. For example, specific combinations of motifs (WRKY, tll, MBP1 and exd) engaged in fine-tuning the MTI response are recognised and targeted for suppression following effector delivery. It is well documented that some WRKYs can act as repressors of MTI depending upon the context of the response (Pandey and Somssich, 2009). Conversely, we found motifs (ABRE, MycN and G-box) enriched in genes rapidly suppressed in response to MAMP recognition yet highly over-represented in two effector-induced clusters. Collectively, these data strongly support the hypothesis that ETS deploys active transcriptional suppression at specific promoter configurations that are targets of MTI.

Our analysis further provided evidence for unique motifs in DC3000-responding gene clusters, identifying specific motifs in both early and late induced clusters and in clusters with genes rapidly suppressed by DC3000 challenge. Rein-

forcing a role for chromatin remodelling in disease progression, we identified both histone OCTAMER and NFY motifs in effector modulated genes. Notably, NFY TFs are implicated in recruiting histone deacetylases to the promoters (Dolfini et al., 2012).

In parallel we used the Wigwams algorithm (Polanski et al., 2014) to identify 31 gene modules with a strong statistical probability of being co-regulated across the multiple time series. Consistent with the over-representation of MYB binding domains (Category F in Fig. 3.4), we identified a module of 17 genes containing a MYB TF binding motif in their promoters, many of which have experimental evidence supporting a role in plant defence. ABA is central to DC3000's virulence strategy (de Torres-Zabala et al., 2007, 2009). Wigwams predicted a module of 26 genes whose promoters are enriched in bZIP and ABI3/VP1 TF binding domains (ABI3/VP1 TFs are regulators of ABA signaling). Wigwams also identified a module of suppressed genes over-represented in photosynthetic components that contain bZIP, bHLH and TCP binding motif in their promoters. Thus these results highlight the utility of using co-regulatory predictions to provide insight into how components of complex processes may be co-ordinately transcriptionally regulated.

3.3.5 Network modelling highlights key effector-modulated genes

Finally, we used CSI to generate a TF network model of the regulatory events controlling transcriptional reprogramming during infection and defence. We focussed on transcriptional events up to 8 hpi, thus capturing the networks recruited during the crucial stages of infection and defence and before the first detectable increase in bacterial growth. The network was unexpectedly sparse, with only 609 interactions predicted between 432 TFs. This may reflect the evolved nature of effector activity, with effectors targeting multiple host proteins to activate parallel, relatively independent signalling pathways rather than a cascading transcriptional response. We identified a core MTI network not perturbed by effectors and interconnected modules with markedly contrasting responses to the presence of effectors. The model provides predictive information about the consequences of activated MTI and possible tactics pathogens use to cause disease. It predicts that a strategy for rapid transcriptional activation of pathways by effectors may be to suppress negative regulators of susceptibility targets, and it is likely these include MYB domain containing TFs. It also predicts that the strong suppression of TFs involved in developmental processes, particularly floral and leaf development, by MAMPs may underlie the mechanism that regulates the trade off between growth and defence. The modelling predicts three TFs, XND1, FBH3, and the bZIP AT2G40620 are at the apex

of a transcriptional cascade regulating a number of the TFs that collaborate to suppress host defences. A future challenge is to identify the specific TFs directly post-translationally modified by effectors (Li et al., 2015) that initiate the transcriptional cascade including how effector induced chromatin changes remodel the nature of the genomic landscape to facilitate the transcriptional reprogramming by early effector induced TFs to permit disease progression.

3.4 Methods

Arabidopsis Sowing And Plant Growth Conditions: *Arabidopsis thaliana* seed (Col-4) were suspended in sterile 0.1% agarose, stratified for 3 days at 4°C in darkness. Seeds were sown using a disposable Pasteur pipette onto ∼7cm square pots of sieved compost mix (Levingtons F2 compost+sand (LEV206): vermiculite (medium grade) mixed in a 6:1 ratio). Plants were grown in individual pots grown in trays under a controlled environment conditions comprising 10 h photoperiod of 120 μmol m^{-2} s^{-1} at 23°C (day), 20°C (night) and relative humidity of 65%. Trays were repositioned every 3-4 days during the growth phase to negate position effects in the growth room. Two days prior to experimentation plants were separated into individual pots and randomized. The challenged leaf, leaf 8, was identified and marked the day before and all plants for each treatment and time point were selected randomly from these to reduce any systematic error.

Pseudomonas growth and maintenance: Maintenance and challenge of bacteria were as described (de Torres et al., 2006). *Pseudomonas syringae* pv. *tomato* strain DC3000 carrying the empty broad host range vector pVSP61 (Innes et al., 1993) and the disarmed DC3000*hrpA*- mutant strain were grown on solidified Kings B media (King et al., 1954) containing rifampicin 50 μg ml^{-1} and kanamycin 25 μg ml^{-1}. For inoculation, overnight cultures were grown at 28°C. Cells were harvested, washed and resuspended in 10 mM $MgCl_2$. Cell density was adjusted to OD_{600} 0.15 (∼0.75 x 10^8 colony forming units (cfu) ml^{-1}).

Pathogen challenges: Treatments were begun 2.5 h after subjective dawn using 34 day old plants by infiltration on the abaxial surface with a 1ml needleless syringe containing bacteria (OD_{600} 0.15; ∼0.75 x 10^8 colony forming units (cfu) ml^{-1}) or 10 mM $MgCl_2$ (mock). Leaf 8 on four individual plants was challenged per treatment (DC3000, DC3000*hrpA*- or mock) and inoculated plants were left under a light bench in the laboratory (22°C). Samples were taken 0, 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 17.5

hpi. Leaves were harvested by cutting at the petiole/leaf blade junction and were immediately snap-frozen and stored at -80°C until used for RNA preparation.

Bacterial growth measurements: All bacterial growth measurements were determined from a minimum of 5 independent replicates, each comprising three challenged leaves/plant. Significant growth differences between treatments were determined by Students t-test ($p < 0.5$), error bars representing standard deviation of the mean. All experiments were repeated at least three times.

Plant RNA extraction: RNA was extracted according to de Torres et al. (2006) from a single leaf ground to a fine powder in a liquid nitrogen pre-cooled mortar. The resultant RNA was cleaned up using a Qiagen RNeasy Plant mini kit according to the manufacturers instructions and samples eluted in 30µl in RNAse free water.

Microarray hybridization: Cy3 and Cy5-labelled cDNA probes were generated from extracted RNA samples, hybridized to CATMA arrays (Allemeersch et al., 2005) and the arrays were processed as described in Breeze et al. (2011). Samples were labeled and hybridized to arrays according to a randomized loop design (as described in Breeze et al. (2011); Fig. S1), facilitating key sample comparisons, whilst minimizing the number of arrays required. Four biological and four technical replicates (including a dye swap) were used for each combination of three treatments and 13 time points; in total, 312 two-arrays were used to generate the transcriptome dataset.

Analysis of microarray data: Data quality checks and normalisation were conducted using a locally-adapted MAANOVA package workflow (Wu et al., 2003). Spatial and dye-bias artifacts were removed from the data through normalisation using locally weighted scatterplot smoothing transformation. Data are deposited at GEO under the accession number GSE56094.

Identification of DEGs: The following selection of methods was used to rigorously capture genes showing differential expression. For Figure 3.1 the Bioconductor package LIMMA (Linear Models for Microarray Data) was applied to $\log_2$ transformed and normalised datasets (Smyth et al., 2005) applying multiple testing correction method separate, a p-value cut-off of 0.05 and false discovery rate correction (FDR) using the Benjamini-Hochberg method. For Figure 3.2 onwards DEGs were selected based upon the three methods described below and probes above the

point at which the false positive level exceeds 10% were included in the list of differentially expressed probes. The three lists of differentially expressed probes for each pairwise comparison of treatments were merged, duplicate probes removed, and the list ordered based on the combined rank from all methods.

MAANOVA (Wu et al., 2003) was used to fit an ANOVA model to the data, from which normalised relative probe expressions were obtained. Per-probe F tests were conducted and probes were ranked using the resulting F statistics. Probes above the point at which the percentage of false positives exceeded 10% were deemed differentially expressed.

A locally-adapted Gaussian Process two-sample (GP2S) method (Stegle et al., 2010) applied Gaussian processes to model the time-series data to infer a likelihood of differential expression. The resulting per-probe Bayes factors were ranked from most to least likely to be differentially expressed.

Bayesian Analysis of Time Series Data (BATS; (Angelini et al., 2008)) adopts a Bayesian approach to estimating expression profiles, identifying genes differentially expressed over time, and to rank them. Ratios of the treatments were used as input expression values.

Clustering by gene expression profile: SplineCluster (Heard et al., 2006) was implemented to cluster DEGs that share similar expression profiles, with sweepmergers between allocations to allow movement of allocated genes between clusters to maximise the resulting log likelihood. Genes for the DC3000*hrpA*- vs. mock comparison were clustered with a prior precision of 0.0001 on the basis of DC3000*hrpA*- expression. DEGs for the DC3000 vs. mock comparison were clustered using DC3000 expression profiles with a prior precision of 0.0001. DEGs between DC3000*hrpA*- and DC3000 were clustered using a prior precision of 0.0005 using the difference in expression between the two infections ($\log_2$DC3000*hrpA*--$\log_2$DC3000). In each case clustering was carried out using a range of prior precision values and the most informative set of clusters selected balancing variation within each cluster v. variation between clusters. Sweepmergers for the clustering of all comparisons was set to 10,000 iterative reclassifications.

Gene Ontology (GO) enrichment: Gene ontology enrichments were assessed using BiNGO (Maere et al., 2005).

Time of first differential expression: An in-house implementation of the Gaussian process gradient analysis (Breeze et al., 2011) was used to identify the time at

which genes first exhibit differential expression, using a delta expression profile for each gene obtained by subtracting expression levels in bacterial challenged leaves from mock expression.

Promoter motif and transcription factor family analysis: MEME was used to search for any enriched motifs within 200bp of the promoters of genes affected early by effectors. Publically available position-specific scoring matrices (PSSMs) were collected from the PLACE and JASPAR databases (Higo et al., 1999; Sandelin et al., 2004). To remove redundancy PSSMs were clustered by similarity, and a representative of each cluster was chosen for screening. Promoter regions corresponding to 200 bp upstream of the transcription start site were from Ensembl Plants database (release 50).

For any given PSSM and promoter the sequence was scanned and a matrix similarity score computed (Kel et al., 2003) at each position on both strands. p-values for each score were computed from a score distribution obtained by applying the PSSM to randomly generated sequences. A binomial test for the occurrence of k sites with observed n values within a sequence of length 200 bp was preformed on the top k non-overlapping hits. The parameter k was optimized within the range 1 to 5 for minimum binomial p-value to allow detection of binding sites without a fixed threshold per binding site. To determine the presence or absence of a PSSM in a promoter, the top 1000 promoters, sorted by p-value, were selected. For each PSSM, its frequency in promoters of each cluster was compared with its occurrence in all promoters in the genome. Clusters were down-sampled (R Stats 'sample' function) to 100 genes, to allow better comparison of hypergeometric p-values across differently sized clusters. Motif enrichment was calculated using the hypergeometric distribution (phyper function in the R stats package). p-values $\leq$ $1e^{-4}$ were considered significant, to allow for multiple testing.

Wigwams module mining: Wigwams (Polanski et al., 2014) was used to identify groups of genes statistically significantly co-expressed across two or more of the three time course datasets. The gene list was filtered to only include genes deemed differentially expressed in at least two of the three performed pairwise comparisons. The gene expression profiles were not standardized, in order to capture the scope of the dynamics in the modeling.

Network modeling: A joint network model for all three treatments was inferred using CSI (Penfold and Wild, 2011). A pre-selection of genes limited the ones used in

modeling to TFs differentially expressed in at least two of the three performed pairwise comparisons, and showcasing a TOFDE of no later than 8 hours in at least one of the performed TOFDE analyses. A pathogen growth profile was also present in the data as a putative network node, with the values for the missing points obtained with a spline fit. The Gaussian process hyperparameter prior was left unchanged and the maximum indegree was set to 2. The marginal probability of each possible regulatory connection was then calculated to generate a marginal adjacency matrix; a threshold probability of 0.1 was used to generate final networks. Connections not joined with the resulting main network structure were trimmed. For the purpose of Figure 3.7 groupings, the area between the expression profile curves for DC3000 and DC3000*hrpA*- was computed for each of the genes and the network was divided into eight sub-network groups linearly spaced between the observed minimum and maximum.

Acknowledgements: This work was supported by the BBRSC funded grant Plant Response to Environmental Stress Arabidopsis Systems Approaches to Biological Research (SABR) programme grants BB/F005806/1 to J.B., K.D., V.B-W., D.R., D.L.W., S.O., C.H., J.P., C.P., D.J., J.M., B.F., L.Bu., P.B. and L.Ba., BB/F005903/1 to W.T., M.T-Z., M.G. and N.S., Leverhulme Trust to M.G. and S.K. and EPSRC grant EP/I036575/1 to C.A.P and D.L.W.. S.J. was funded by an Exeter Systems Biology studentship. R.H. was funded by the Engineering and Physical Sciences Research Council (EPSRC)/BBSRC funded Warwick Systems Biology Doctoral Training Centre; L.L. was funded by a BBSRC SABR studentship.

Author Contributions: L.L., W.T., M.T-Z., G.L. and M.G. carried out experiments. W.T., M.T-Z., L.L., K.P., S.J., L.B., J.M., C.A.P, D.J.J., C.H., L.B., J.P., A.M., J.S., R.H., D.R., D.L.W., S.O., V.B-W., N.S., K.D., J.B., M.G. were responsible for data processing, analysis and tool development. L.L., K.P., S.J., A.M., L.B., D.L.W., D.R., J.B., K.D., S.O., V.B-W., M.G. and K.D. all had input into the design of the experiments and analysis. L.L., K.D. and M.G. wrote the manuscript with contributions from all authors.

Chapter 4

Inference of functional gene regulatory networks mediating Arabidopsis response to environmental stress

The work within this chapter will be shortly submitted to the Plant Physiology journal. I am first author. In the event of publication, the supplementary data will be available online on the journal's site; it can currently be accessed at `http://www2.warwick.ac.uk/fac/sci/sbdtc/people/students/2011/krzysztof_polanski/thesis_supplement/` (password: 21datasets).

The inference of causal regulatory network models from expression data is a very computationally intensive task, and it is quite common to see the number of profiles used in the inference scaled down through gene selection (Penfold and Wild, 2011) or clustering (Windram et al., 2012). This work is based on transcription factor-only regulatory network models inferred by Dr Christopher Penfold, showing *Arabidopsis thaliana* response to *Botrytis cinerea* and *Pseudomonas syringae* pv. *tomato* DC3000 infection, drought, high light, and long and short day senescence. The motivation behind the work was an attempt at introducing downstream target information into these regulatory networks. Models inferred on transcription factors only are quite informative due to them capturing the main regulatory interactions happening in the organism, but lack downstream target information that would make it possible to assess the actual role of the transcription factors in regulating specific physiological responses. By augmenting the networks with groups of genes putatively co-regulated across multiple stress responses identified

by Wigwams, the innate tendency for regulatory networks to yield focused, easily experimentally testable hypotheses can be used to full effect to propose potential key regulators for further investigation. This was achieved through GO term overrepresentation analysis of the immediate downstream targets of network nodes, with a focused examination of the *B. cinerea* and *P. syringae* infection response networks revealing the involvement of multiple known and novel transcription factors in jasmonic acid, salicylic acid, abscisic acid and ethylene signalling.

For this study, I created the algorithms, conducted the analyses, interpreted the results, wrote the manuscript, compiled the supplementary data, and made figures under the guidance of Dr Katherine Denby. The exceptions are the single transcription factor binding motif overrepresentation analysis, proposed by Dr Richard Hickman (with subsequent reimplementation by me), and M-VBSSM modelling created and performed by Dr Christopher Penfold, with his notes and materials being used for the creation of Figure 4.1 and the relevant parts of the manuscript.

Inference of functional gene regulatory networks mediating Arabidopsis response to environmental stress

Krzysztof Polanski[1], Christopher A. Penfold[1,§], Richard Hickman[2], Murray Grant[3], Phil Mullineaux[4], Uli Bechtold[4], David Rand[1], Jim Beynon[1,5], Vicky Buchanan-Wollaston[1,5], Sascha Ott[1], David L. Wild[1] and Katherine J. Denby[1,5]

[1] Warwick Systems Biology Centre, University of Warwick, CV4 7AL, UK

[2] Department of Plant-Microbe Interactions, Utrecht University, 3584 CH, Utrecht, The Netherlands

[3] Biosciences, College of Life and Environmental Sciences, University of Exeter, EX4 4QD, UK

[4] School of Biological Sciences, University of Essex, CO4 3SQ, UK

[5] School of Life Sciences, University of Warwick, CV4 7AL, UK

Current Address:

[§] The Wellcome Trust/CRUK Gurdon Institute, University of Cambridge, CB2 1QN, UK

Corresponding author: Dr Katherine Denby, k.j.denby@warwick.ac.uk

Short title: Network modelling of Arabidopsis stress response

Keywords: Gene regulatory networks, plant stress, transcriptomics, network inference, Arabidopsis, biotic and abiotic stress, senescence

Whilst the response of a given organism to an outside stimulus is a highly complex and hard to quantify phenomenon, the easily measurable transcriptome change provides insight into the regulatory interactions between transcription factors and their targets. If the performed experiment is a time series, inference algorithms can be used to elucidate network models of causal interactions between genes used in the modelling. In this study, we have inferred causal transcription factor-only networks for multiple time series transcriptome data sets of *Arabidopsis thaliana* response to six different biotic and abiotic conditions through M-VBSSM, an expansion of VBSSM that samples a larger gene space for local models subsequently merged into a final joint network. The effect of these networks on the data across multiple conditions was assessed through the identification of co-regulated gene modules with Wigwams. Whilst conventional methods of regulatory assessment, such as transcription factor binding site overrepresentation analysis, provided some information on the underlying regulatory interactions, more understanding of the responses was obtained by using the Wigwams modules to expand the M-VBSSM TF-only models with potential downstream genes. These enhanced network models offer a high degree of functional insight, and the *Botrytis cinerea* and *Pseudomonas syringae* infection response networks were analysed for primary defence hormone signalling nodes, revealing a number of both known and novel genes potentially involved in this aspect of the response. The analysis of the intersection of these two network models revealed a novel five-gene interaction, which may be involved in defence hormone signalling across both of the responses.

4.1 Introduction

Transcriptional reprogramming is a major part of a plant's response to internal and external stimuli, with a large number of genes typically changing in expression under any given condition (for example, Breeze et al. (2011), Windram et al. (2012), Lewis et al. (in press)). The regulatory mechanisms underlying this transcriptional re-programming are not merely linear signalling pathways, but rather more intricate regulatory networks, with interactions between modules driving different responses (Shinozaki et al., 2003). Such networks are predominantly driven by transcription factors (TFs), with non-TF genes being key downstream targets that carry out physiological and/or metabolic functionality. The ability to elucidate and model these transcriptional networks is vital to comprehending and manipulating plant stress responses. It enables the prediction of key target genes in the network to enhance stress resilience.

There are both experimental and computational approaches that enable the identification of regulatory interactions between TFs and their targets. ChIP-seq (Park, 2009) can identify binding sites of a particular TF throughout the genome, whereas yeast one-hybrid (Deplancke et al., 2004) can identify TFs that bind a specific promoter. However, performing these assays on a genome-wide scale (i.e. for all TFs and targets in the genome) would be prohibitively time consuming, making computational network inference a valuable preliminary analysis to pinpoint relevant promoters and TFs for focused experiments.

The type of network model that can be generated depends on the available data (Windram et al., 2014). In the case of a large number of multiple static datasets, the most common type of data available, connections can be created between genes based on the correlation of their expression across conditions. The resulting co-expression networks can be mined using a guilt-by-association approach to predict function of unknown genes based on their proximity to genes of known function (Usadel et al., 2009). Such network models lack causality, as there is no way to predict which gene is responsible for the observed expression profiles. Time series data provides more insight into the underlying dynamics of transcriptional reprogramming, making it possible to produce network models with directed edges. Comparative testing of multiple inference algorithms on DREAM4 data (Penfold and Wild, 2011) identified CSI (Klemm, 2008; Penfold and Wild, 2011) as the preferable method, but noted that its computational tractability scales poorly with the number of profiles involved. VBSSM (Beal et al., 2005) performed adequately in the trials, and offers an extra degree of flexibility with incorporation of hidden states. This makes the

model capable of accounting for missing time-point observations and variables, such as genes not measured or non-transcriptional processes. Both of these approaches produce causal transcriptional network models of the regulation underlying changes in gene expression.

These time series network inference algorithms are computationally expensive and the number of expression profiles used as input needs to be limited. Whilst analyses have been performed on representative profiles of co-expressed gene clusters (Windram et al., 2012), performing inference on individual genes provides more information on signal flow. TF expression determines the network topology and hence modelling with differentially expressed TFs only provides a good balance between information and tractability (Lewis et al., in press). However, for network models to predict function it is essential that networks include non-TF encoding genes that control the actual physiological response of the plant.

In this paper we infer TF-only network models capturing the regulatory events underlying *Arabidopsis thaliana* response to pathogen infection (*Botrytis cinerea* and *Pseudomonas syringae*), drought, high light, and long- and short-day senescence. We present a novel strategy for expanding these networks with non-TF genes using an algorithm that predicts co-regulation (Polanski et al., 2014). These augmented network models are used to predict TFs responsible for controlling specific biological processes. Hence we provide functional network models of key plant stress responses that can be used by the community to drive further experimental testing.

4.2 Results

4.2.1 Elucidating TF network models underlying Arabidopsis responses to abiotic and biotic stress

Our network inference approach depends on the availability of time series transcriptome data sets. We used 6 such data sets all obtained from Arabidopsis leaves and generated using the same technical platform. These data sets included the transcriptome response to infection with the fungal pathogen *Botrytis cinerea* (Windram et al., 2012), infection with the bacterial pathogen *Pseudomonas syringae* pv *tomato* DC3000 (Lewis et al., in press), drought (Bechtold et al., in preparation), high light (Alvarez-Fernandez et al., in preparation) and long- and short-day natural senescence (Breeze et al., 2011). The number of time points ranged from 13 to 24 (not necessarily linearly spaced) with four biological replicates at each time point. The biotic and abiotic stress experiments (*B. cinerea*, *P. syringae*, drought and high

light) also included controls at each time point which were used to identify genes differentially expressed specifically in response to the stress. For the senescence time series, genes differentially expressed over time were determined. Details of the time series data sets and differentially expressed genes are provided in the relevant publication or in the methods section.

The temporal nature of these Arabidopsis transcriptome data sets, along with the high resolution and replication, facilitates network inference to predict causal regulatory relationships between individual genes (Windram et al., 2014). However, thousands of genes are differentially expressed in each time series and given the data structure (i.e. number of genes, time points and replicates) network models generated using such a large number of genes are unlikely to be better than random. We took two approaches to overcome this. Firstly, we modelled only differentially expressed genes encoding transcription factors (TFs) (Supplementary Dataset 1) as, in transcriptional networks, these regulatory genes determine the network topology and information flow. Secondly, we generated a Metropolis wrapper for the Variational Bayesian State Space Modelling (M-VBSSM) algorithm (Beal et al., 2005) that enables a probabilistic search for local network optima.

The M-VBSSM network inference algorithm is illustrated in Figure 4.1. VBSSM is a network inference algorithm utilising steady state modelling, with hidden states making it possible to account for regulatory interactions not captured in the experiment measurements. VBSSM, like most network inference algorithms, is computationally intensive, and direct inference of an underlying network model spanning hundreds of TFs is intractable. Instead a consensus network was created by combining VBSSM networks inferred from a smaller group of TF genes. Optimal gene selection for these smaller network models is a nontrivial task. A network model was initiated with a set of 79 randomly chosen TF genes and the seed gene (Figure 4.1A). After inference of a network model, a randomly chosen subset of the TF genes (omitting the seed gene) were switched for the same number of randomly chosen genes that were not used in the previous modelling (Figure 4.1B). This switching is performed in a probabilistic manner, in that as the network models improve fewer genes are switched. The switching and network inference is repeated for 2000 iterations, yielding a locally optimal network model (Figure 4.1C). Network models from individual iterations are discriminated on the basis of their marginal likelihoods, with larger values representing superior models, and this value reaches convergence during the 2000 iterations of the algorithm (Figure 4.1D). The entire network inference procedure is then repeated with a different seed gene (and subsequently for every gene in the modelling) (Figure 4.1E). The individual locally optimal networks

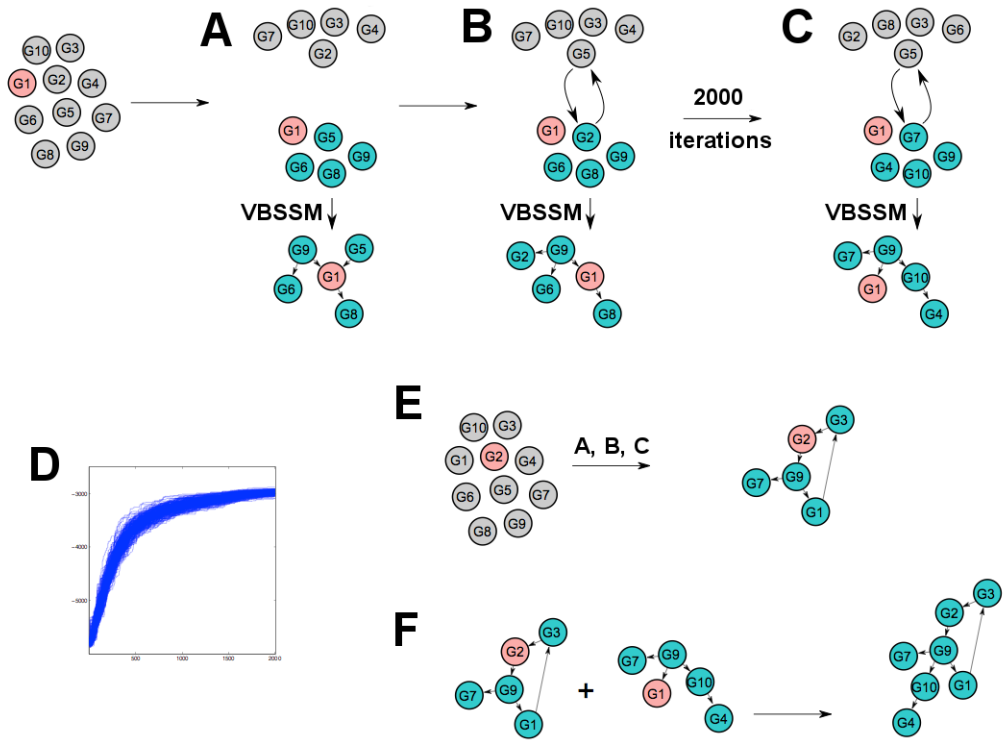

Figure 4.1: The process of obtaining M-VBSSM network models. Focusing on gene G1, a fixed size gene set is obtained from the pool of all genes available in the modelling and VBSSM inference is performed **(A)**. Then, N genes, where N is binomially distributed, are switched at random between the selected and unselected genes without the possibility of removing G1, and VBSSM is used to obtain an alternate local model **(B)**. This process is repeated 2000 times **(C)**, yielding the final local network model for G1. Individual VBSSM models are judged based on their marginal likelihood, with a Metropolis sampler yielding a likelihood plateau in the obtained fit quality **(D)**. The same procedure is repeated for all other genes in the set **(E)**, and a final network model is obtained as a union of all the local networks centred on each individual gene **(F)**.

were then combined into a single consensus model (Figure 4.1F). In this way we generated TF network models for each response from the TFs differentially expressed in each time series data set.

The network models for the TF-TF regulatory interactions mediating each response are available in Supplementary Dataset 2. In these networks each node is a TF gene and an edge indicates a regulatory influence from one TF to another, with the frequency (number of models the edge occurred in) and probability estimate (Z score) of each edge given. A Z-score threshold of 1.65 and an edge frequency threshold of 1 were applied.

These network models predict the transcriptional regulation between TFs that drive the large changes in gene expression in response to biotic or abiotic stress or natural senescence. Using the frequency and/or probability associated with each regulatory interaction (edge), higher or lower stringency networks can be extracted, and the highest confidence regulatory interactions identified. Network features (for example, connectedness) can be used to predict TFs with a key role in each biological response.

4.2.2 Identifying regulatory footprints in transcriptome data

Gene co-expression is often used as an indication of co-regulation, and co-expression of a number of genes involved in the same pathway suggests specific activation or repression of this pathway. We used the Wigwams algorithm (Polanski et al., 2014) to identify groups of genes co-expressed across the multiple time series data sets. Wigwams can identify statistically significant co-expression (rather than that arising from the abundance of a particular expression profile) and such gene modules co-expressed across multiple time series are likely to be co-regulated. Downstream target co-regulation events, such as those detected by Wigwams, stem from the action of upstream transcription factors, and can be termed regulatory footprints.

Genes that are differentially expressed in two or more of the six time series data sets were used in the Wigwams algorithm. This identified 71 gene co-expressed gene modules spanning two to five of the time series and containing 5434 unique genes (Supplementary datasets 3 and 4). The data set combinations, and number of modules, identified are shown in Figure 4.2, with the total number of modules spanning each pair of conditions given in Table 4.1. Firstly, it is apparent that Wigwams identified co-expressed gene modules across many different data set combinations. No genes were found to be significantly co-expressed in all 6 time series but 3 modules were identified spanning 5 time series data sets (two spanning biotic stress, senescence and drought, and one module spanning biotic stress, senescence

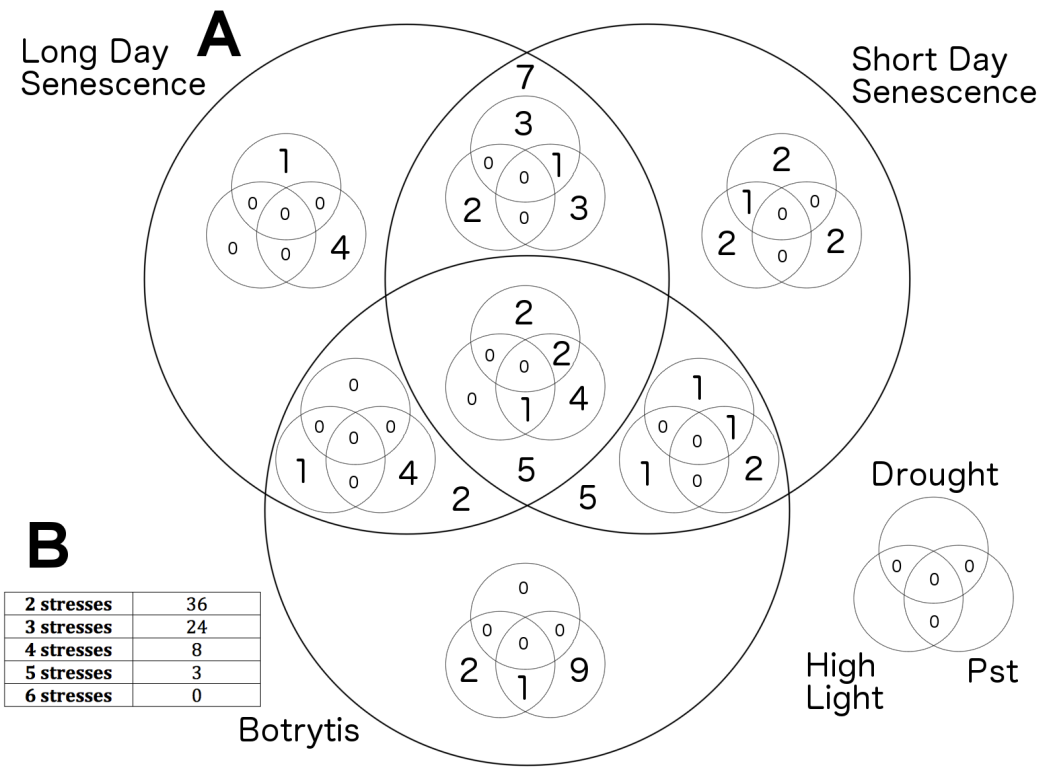

Figure 4.2: The number of Wigwams modules identified for each condition combination **(A)** and the distribution of module totals across the number of conditions they span **(B)**. The individual condition combinations are represented through shading of the appropriate columns. The highest number of Wigwams modules spanning any individual condition combination is 9, for *B. cinerea* and *P. syringae* infection responses. In terms of numbers of conditions spanned by the modules, the highest number of modules (36) feature potential co-regulation across a pair of conditions, with the numbers going down as the number of conditions spanned goes up. No modules were identified for all six conditions.

Table 4.1: The distribution of Wigwams modules spanning each pair of conditions

Conditions	Drought	High Light	*P. syringae*	LD Sen	SD Sen
B. cinerea	6	6	24	21	24
Drought		1	4	9	13
High Light			2	4	7
P. syringae				19	16
LD Sen					30

The totals shown in the table include modules spanning more than two conditions — for example, if a module spans three conditions, it will be included in the totals for all three of the condition pairs. LD Sen - long day senescence; SD Sen - short day senescence

and high light). The variety of time series combinations included in gene modules suggests that rather than overall transcriptional responses being conserved between different stresses, specific regulatory modules (regulator and target genes) have been co-opted to different stress responses. The number of modules including a particular condition is not strictly correlated to the number of genes identified as differentially expressed in that time series, for example the high light data set has 6703 differentially expressed genes and is part of 11 modules, whilst short day senescence with 6241 differentially expressed genes is part of 47 modules (Figure 4.2).

As expected the two time series from natural senescence (long day and short day) show the highest number (30) of co-expressed gene modules. There is a higher number of shared modules among the two senescence and pathogen infection responses than with, or between, the two abiotic stress responses. In the case of drought, this may reflect the relatively low number of differentially expressed genes, but it is also likely to reflect the difference between the transcriptional responses to biotic and abiotic stress. This difference is not just a difference in direction of differential expression (as co-expressed gene modules do not need to change in expression in the same direction in each stress) but a difference in the identity or regulation of differentially expressed genes. The highest number of unique modules, i.e. just co-expressed between those two time series, was observed for *B. cinerea* and *P. syringae* infection (Figure 4.2) suggesting a number of biotic stress-specific discernible regulatory modules.

We would expect co-regulated gene modules to be enriched for genes involved in the same biological process. We assessed the 71 Wigwams modules for overrepresentation of Gene Ontology (GO) (Ashburner et al., 2000) terms using BiNGO (Maere et al., 2005). 59 of the 71 Wigwams modules contained genes enriched for at

least one GO term annotation (Figure 4.3, Supplementary Dataset 5) suggesting a common function. Response-related GO terms appear to be overrepresented, both for immediately relevant (response to chitin in modules spanning *B. cinerea* infection) and unrelated (response to salt in a module spanning *P. syringae* infection and long day senescence) functionality. The unrelated overrepresented responses hint at crosstalk between networks regulating response to different stimuli (Shinozaki et al., 2003). A high number of modules feature genes related to chloroplasts and ribosomes, with the general expression trend for those modules being down-regulation. Additionally, module 8 (spanning drought, long and short day senescence) features eight genes responsible for chromatin assembly and nucleosome organisation (P-value $\sim 10^{-8}$). This is in line with Lewis et al. (in press) suggesting that the down-regulation of nucleosome genes may happen in response to other stimuli as well. In contrast to the uniformly down-regulated modules' GO terms relating to photosynthesis, chloroplasts and ribosomes, the GO terms of the uniformly up-regulated modules relate to a variety of biosynthetic, metabolic and catabolic processes, as well as autophagy. This suggests an alteration of signalling and metabolism to adapt to the external stimuli, conserved across responses to different environmental factors. Module 6 (spanning *B. cinerea* and *P. syringae* infection) features an over-representation of oxidoreductase activity (P-value $\sim 10^{-7}$), with specific processes involving organic acid catabolism (P-value $\sim 10^{-5}$) and auxin metabolism (P-value $\sim 10^{-4}$). In addition, the uniformly up-regulated modules are the only module group to feature GO terms related to the four primary defence response hormones — five modules have ABA-related GO terms, two have JA-related GO terms and one has an ethylene-related GO term. 14 of the 71 modules lack a uniform expression trend, with the genes being up-regulated in some of the spanned conditions and down-regulated in others. The functionality of these modules is the least clearly defined, with 4 of them (the highest proportion of the regulatory groups) lacking any GO terms, and the identified GO terms showing overall feebler P-values than those appearing in the uniformly up- or down-regulated modules. The overall functionality of these modules was in line with the down-regulated ones, with GO terms related to chloroplasts and photosynthesis being prevalent. It should be noted that the condition where the module would be up-regulated was usually high light, which would logically indicate the induction of expression of choloroplast/photosynthesis genes by the stimulus whilst being repressed by other conditions.

Genes co-expressed across multiple time series data sets, and with shared biological functions, have a high likelihood of being co-regulated. We looked for evidence of co-regulation of Wigwams modules by testing for overrepresentation of

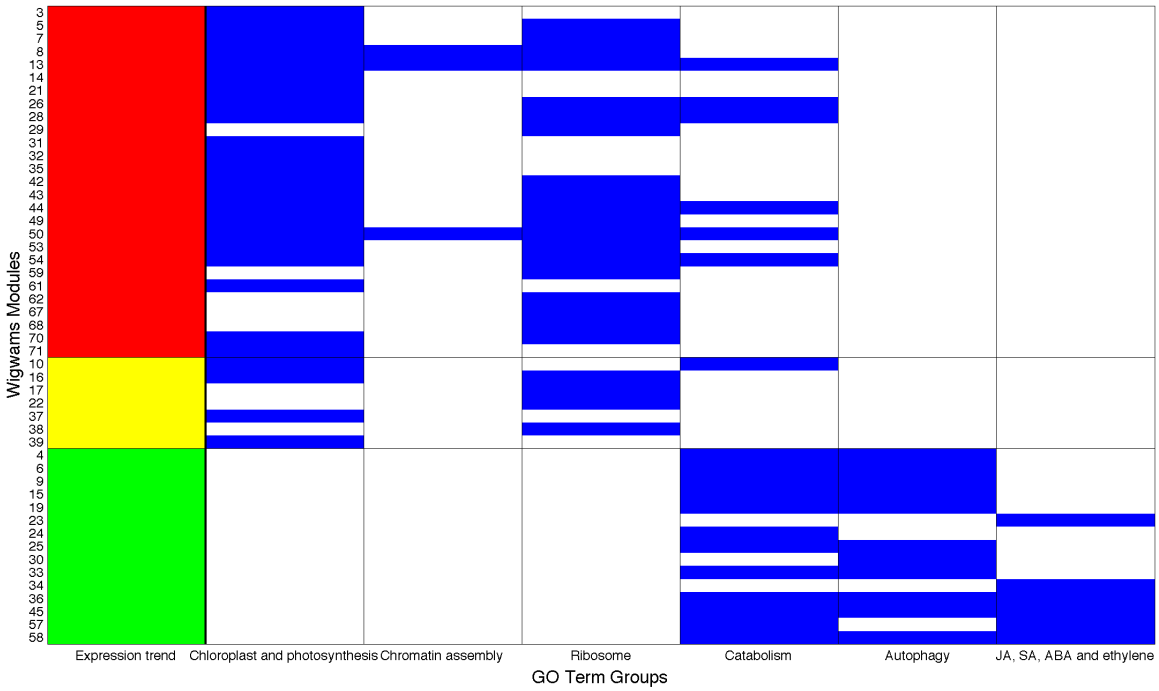

Figure 4.3: The overrepresentation of GO term groups among gene members of Wigwams modules. The Wigwams modules in the figure are grouped on expression trend, as marked in the first column of the heatmap, with red indicating down-regulation across all of the conditions spanned by the module, green indicating up-regulation across all of the conditions spanned by the module, and yellow indicating a lack of regulatory homogeneity, with some of the conditions spanned by the module featuring up-regulation and some featuring down-regulation. The presence of an overrepresented GO term matching the group is marked in blue.

444 known TF binding motifs in the genes' upstream promoter sequences. These motifs were identified via protein binding microarrays and analysis of DNA binding domains (Franco-Zorrilla et al., 2014; Weirauch et al., 2014). 24 of the 71 modules have at least one known TF binding site overrepresented among the promoters of their gene members (Figure 4.4, Supplementary Dataset 6).

One of the identified modules is a 464-gene group statistically significantly co-expressed across *P. syringae* infection and long day senescence, and the expression profiles of the genes across all six of the time course datasets can be seen in Figure 4.5A, with gene membership as module 36 in Supplementary Dataset 1. The GO terms overrepresented for this module include response to jasmonic acid stimulus and response to abscisic acid stimulus, identifying a shared network footprint between the two conditions. Among the module's genes are ERF1 and MYB108, previously shown to be involved in JA signalling in plant defence (Lorenzo et al., 2003; Mengiste et al., 2003). AT5G59220 was shown to bind to the promoter of SRK2E, a kinase activated by ABA (Umezawa et al., 2009), continuing the hormone response signalling. A closer inspection of the promoters of the genes in the module reveals the overrepresentation of a number of known transcription factor binding sites — M2345, which is bound by PIL5 (AT2G20180) from the bHLH family; M0261 and M0272, which are bound by ABI5 (AT2G36270) and bZIP63 (AT5G28770) respectively, with both of the TFs being members of the bZIP family, and M1408 (bound by ANAC100 (AT3G15170) and CUC1 (AT5G61430)) and M1410 (bound by ANAC058 (AT3G18400)) (Weirauch et al., 2014). The presence of transcription factor binding sites bound by a variety of families suggests the potential for combinatorial regulatory action in the preserved crosstalk between responses to *P. syringae* infection and long day senescence.

The transcription factor binding motifs used for the analysis are precise, with the majority of the sequences being assigned to a single regulator. However, not every overrepresented transcription binding motif has to correspond to a regulatory interaction, as it is possible that only a small number of transcription factors capable of regulating the downstream genes are performing that role in a given stimulus response. On the basis of the overrepresentation predictions, the time series data allows for the identification of exact TFs that may be driving the observed responses. The data for the conditions the module spans can be scanned for TFs that begin to change in expression no later than the potential target genes. The Gaussian process gradient approach (Breeze et al., 2011) was used to identify the first time point at which the genes in the module, as well as transcription factors with known binding sites overrepresented in the promoters of the module mem-

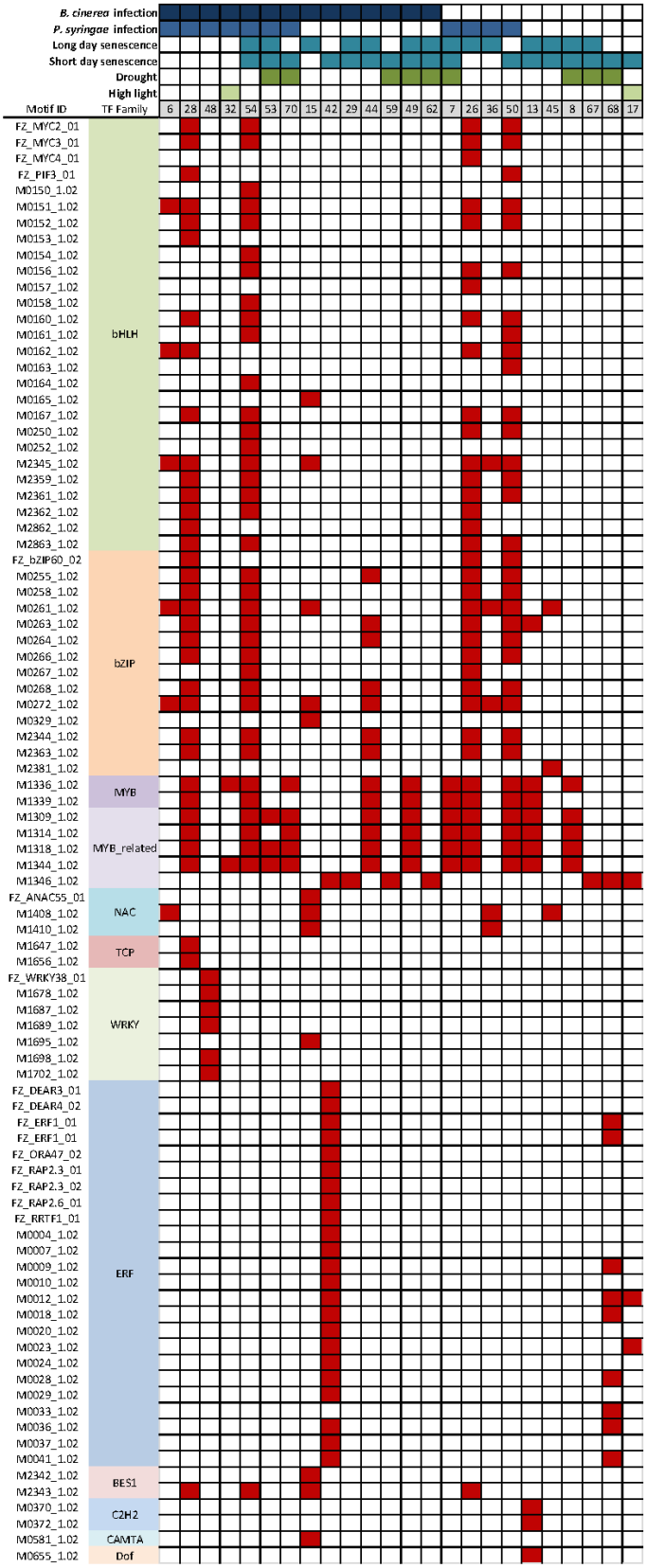

Figure 4.4: The overrepresentation of regulatory motifs bound by known transcription factors in the promoters of genes identified as part of Wigwams modules. The motifs correspond to single transcription factors, as identified by protein binding microarray studies, and the overrepresented binding sites correspond to members of bHLH, bZIP, MYB, MYB-related, NAC, TCP, WRKY, ERF, BES1, C2H2, CAMTA and Dof transcription factor families. It is common for Wigwams modules to have motifs for different transcription factor families overrepresented in the promoters of the member genes, potentially implying a combinatorial regulatory role of transcription factors from many families in the regulation of Wigwams modules across multiple conditions.

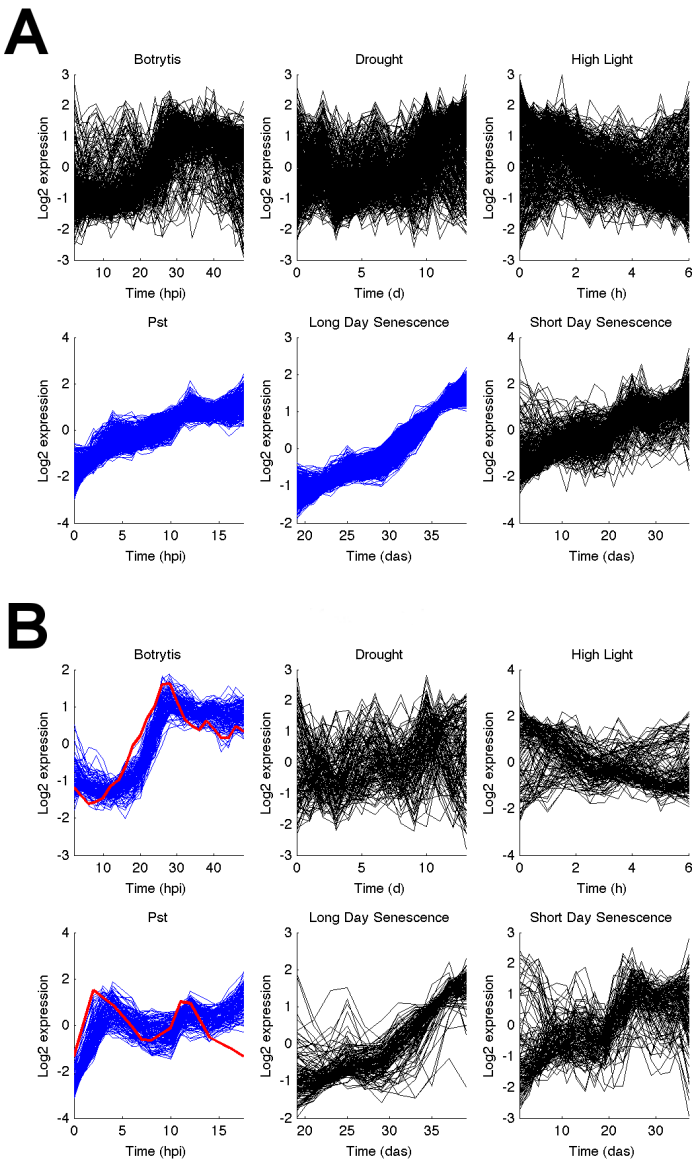

Figure 4.5: **(A)** The expression of a 464-gene Wigwams module spanning P. syringae infection and long day senescence across all of the datasets. This particular module had overrepresented binding motifs for bHLH, bZIP and NAC TFs in its members' promoters, but none of them were identified as putative regulators by the gradient tool follow up analysis. **(B)** The expression of a 136-gene Wigwams module spanning *B. cinerea* and *P. syringae* infection, with the follow-up analysis identifying WRKY33 (expression profile in *B. cinerea* and *P. syringae* overlain over module expression in red) as a putative regulator across both conditions, with its expression profile clearly foreshadowing the module's expression profile. The module's condition span has its expression plotted in blue. The conditions where the expression profiles are plotted in black were not identified as part of the module condition span, and appear far less correlated.

bers, show a discernibly nonzero change in expression profile in all of the datasets the regulated module spans. When applied to the previously identified overrepresented known transcription factor binding sites, this allows for the identification of single regulators that can be driving the expression of the module genes across all of the relevant conditions. A single change time point was defined for each of the module's condition span, and a subsequent comparison of the module gradient with the individual transcription factor gradient allows for the identification of putative regulators. Applying this approach to the genes and putative upstream regulators of module 48, spanning *B. cinerea* and *P. syringae* infection, allows for the identification of WRKY33 (AT2G38470) as a potential regulator of the genes in the module, as it was the only one of the putative WRKYs identified by the overrepresentation analysis to also fulfil the gradient change criteria. The expression profile of WRKY33, along with the expression profile of module 48, can be seen in Figure 4.5B. The visualisation shows WRKY33, highlighted in red, clearly foreshadowing the module expression trend at earlier time points. This is consistent with prior knowledge, as WRKY33 has been shown to be a key element in the regulation of the antagonism between the jasmonic acid response, the course of action for necrotrophic pathogens such as *B. cinerea*, and salicylic acid response, the course of action for (hemi-)biotrophic pathogens such as *P. syringae* (Birkenbihl et al., 2012). Another interesting example comes in the case of module 17, which showcases genes exhibiting alternate regulation in response to high light and short day senescence. The gradient follow-up analysis reveals the putative regulatory role of ERF7 (AT3G20310), ERF8 (AT1G53170) and TBP3 (AT5G67580) in a shared element of the crosstalk between the responses to these conditions being utilised in alternate ways, with up-regulation in high light and down-regulation in short day senescence. The analysis was carried out for every instance of a known transcription factor binding site being overrepresented among members of a Wigwams module, and all of the cases where the expression data revealed the corresponding transcription factor to be a likely putative regulator can be seen in Supplementary Dataset 7. However, this method is limited by the current knowledge of motif binding specificity — in spite of a number of different transcription factor families being predicted to regulate genes within module 36, the gradient analysis did not indicate any of them as putative regulators. As such, it is possible that the motifs are bound by different members of the family, or different transcription factors altogether or that the overrepresented motif is not actually the one driving the analysed expression pattern. In total, of 381 assessed regulatory hypotheses (instances of a transcription factor binding site overrepresented in a module) spread across 24 modules, only 36 such

hypotheses spread across 16 modules had the Gaussian process gradient tool's output solidify the regulatory hypothesis, making the select few predicted transcription factor-module interactions prime targets for experimental validation.

4.2.3 Expanding the transcription factor-only network models

Whilst the transcription factor-only network models provide good insight into the information flow in response to the stimuli, they fail to fully capture the functionality of the resulting transcriptional changes. By integrating Wigwams modules with the TF only network we can extend the network to non-TF genes and make predictions about the function and/or role of individual TFs.

The process of TF-only network expansion with genes contained in Wigwams modules is shown in Figure 4.6. Initial TF-only networks (Figure 4.6A) capture a good degree of regulatory flow, but are lacking in downstream genes carrying out the functionality induced by the regulatory flow. A Wigwams module contains genes likely to be co-regulated, and are rich in downstream targets. Hence if a Wigwams module (the blue nodes in Figure 4.6B) contains multiple TFs also predicted to be co-regulated in the network model (the non-translucent blue nodes), then the assumption can be made that the other (non-TF) genes in this module may also be regulated by the same upstream TF (yellow). For example, in Figure 4.6B four TFs in the module are predicted to be co-regulated by a single network node, hence the whole module is added to the network. Reflecting the aim of Wigwams, and ensuring higher stringency, a Wigwams module is only added to the network if there are four or more TFs predicted to be co-regulated in a single stress (that is significant in the module) or if one or more TFs are predicted to be regulated by the same upstream TF across all conditions significant in the module. The four target TF limitation is in place to preserve and emphasise the underlying regulatory structure whilst only adding the downstream genes we are most confident about. To ensure high stringency of co-regulation, instead of the 71 Wigwams modules produced after merging of similar modules and removal of modules spanning subsets of conditions of other modules (Polanski et al., 2014), the raw initial modules were used. These are smaller and represent the highest statistically significant co-expression groups. Crucially, the newly created connections (Figure 4.6C) are not allowed to influence the addition of other Wigwams modules to the network; only genes within the initial network model are used in deciding whether to extend the network. The resulting expanded network models for each individual condition can be found in Supplementary Dataset 8.

The summary for the size of the network models (number of connections

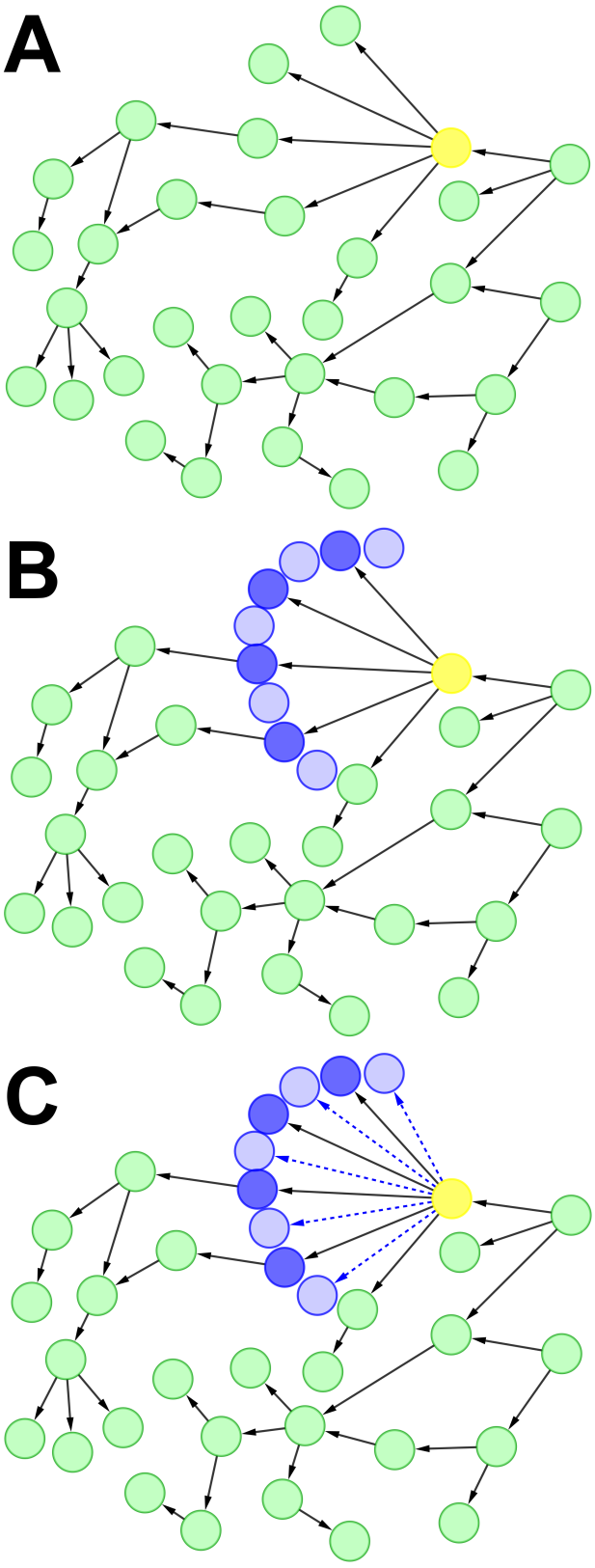

Figure 4.6: The primary process of expanding M-VBSSM networks with non-transcription factor genes using Wigwams modules. The original, transcription factor-only model **(A)** is used throughout the analysis to match Wigwams modules and add extra genes to the network without letting the new connections impact the expansion process. Attempts are made to match each eligible node featuring downstream targets with each Wigwams module spanning the condition the M-VBSSM network was inferred for. In order for a match to be made between the node (in yellow) and a Wigwams module (nodes highlighted in blue), the module genes and the node's downstream targets need to have a sufficiently large membership overlap (solid blue nodes, **B**). The identified connections (marked in blue dashed lines) are subsequently stored to be added to the expanded network model **(C)**.

Table 4.2: The comparison of the number of network edges in the original transcription factor-only M-VBSSM models with the networks expanded with Wigwams modules

Condition	M-VBSSM TF-only model		Expanded network model	
	Edges	Nodes	Edges	Nodes
B. cinerea	8998	526	43919	2623
Drought	1869	162	3752	383
High light	7236	454	7236	454
P. syringae	16058	573	38636	1837
LD senescence	10208	501	61304	3438
SD senescence	1742	256	12252	1614

The most sizeable increase occurs in the short day senescence model, which goes from 1742 edges to 12252 edges. No additional connections are added to the high light model.

and number of nodes) before and after this expansion procedure can be seen in Table 4.2. The highest number of additional connections is created in the long day senescence network, whilst the highest ratio of expanded network connection total to original network connection total is short day senescence. The high light network model does not get expanded with a single additional connection, which could stem from the different regulatory interactions captured by the M-VBSSM model and Wigwams modules, the scarcity of Wigwams modules spanning high light, as well as the potential for smaller numbers of transcription factors acting in unison in high light (with a total of 4 showing up in a Wigwams module being required to expand the network).

The expansion procedure outlined above results in the enhancement of TF-only networks with downstream gene information, which in turn captures information on the actual functionality carried out by the regulatory signalling core. These expanded causal models can be subsequently mined for relevant functionality, allowing for the formulation of succinct experimental hypotheses and helping transition from genotype to phenotype.

4.2.4 Functional inference for TFs mediating *B. cinerea* and *P. syringae* defence response

We were particularly interested in predicting the function of TFs involved in regulating the Arabidopsis defence response against *B. cinerea* and/or *P. syringae*. This combination of stimuli also appears to produce a broad range of different Wigwams modules, as seen in Table 4.1 and Figure 4.2 (including the highest number of

modules exclusively spanning a pair of conditions), indicating a diverse set of gene expression profiles in the network footprint. Both the expanded individual pathogen infection networks and their intersection were chosen for further analysis, with the aim of finding genes with roles specific to a single pathogen and genes with common roles in both responses. It should be noted that the intersection network does not necessarily reflect identical signal flow in both pathogen infections, as the intersection only makes up a small fraction of the original single-pathogen networks and may be playing largely different roles under each stimulus. For example, network modules may contain genes up-regulated during *B. cinerea* infection and down-regulated during *P. syringae* infection — but the regulatory connections between the genes are preserved. A representation of the intersection of the expanded *B. cinerea* and *P. syringae* networks can be seen in Figure 4.7A, and the network files featuring the connections in both the individual infection response models are included as part of Supplementary Dataset 8.

The obtained expanded models feature a high number of nodes and are very interconnected, as previously outlined in Table 4.2. The intersection of the individual defence response networks, as shown in Figure 4.7A, is still very heavily connected, and manual analysis of the connections would be a very lengthy process. Given the causal structure of the network and the presence of a number of downstream genes added during the expansion procedure, automated functionality mining can be performed. To assess the role of a network TF node in defence, we selected the nodes immediately downstream of the TF and assessed GO term overrepresentation within these genes as shown in Figure 4.7B. We tested overrepresentation using the hypergeometric test and maximised the stringency of this analysis by testing overrepresentation in the test set compared to the set of genes that make up all the downstream nodes in the network (Fig. 4.6B). This means the results provide information on the role of the TF of interest in the context of the defence response as a whole. For example, if a large proportion of the network nodes contained GO annotations related to jasmonic acid, testing each downstream group against the genome as a whole would result in a large number of TFs having a potential function in JA signalling. By performing the more stringent analysis, TFs that still show overrepresentation of GO terms related to jasmonic acid are likely to be key regulators of this signalling process. This analysis was carried out both on the individual *B. cinerea* and *P. syringae* infection networks, as well as the higher confidence intersection network. The number of TF nodes with downstream targets (and hence eligible for downstream target GO term overrepresentation testing), as well as the number that have at least one GO term overrepresented in their target

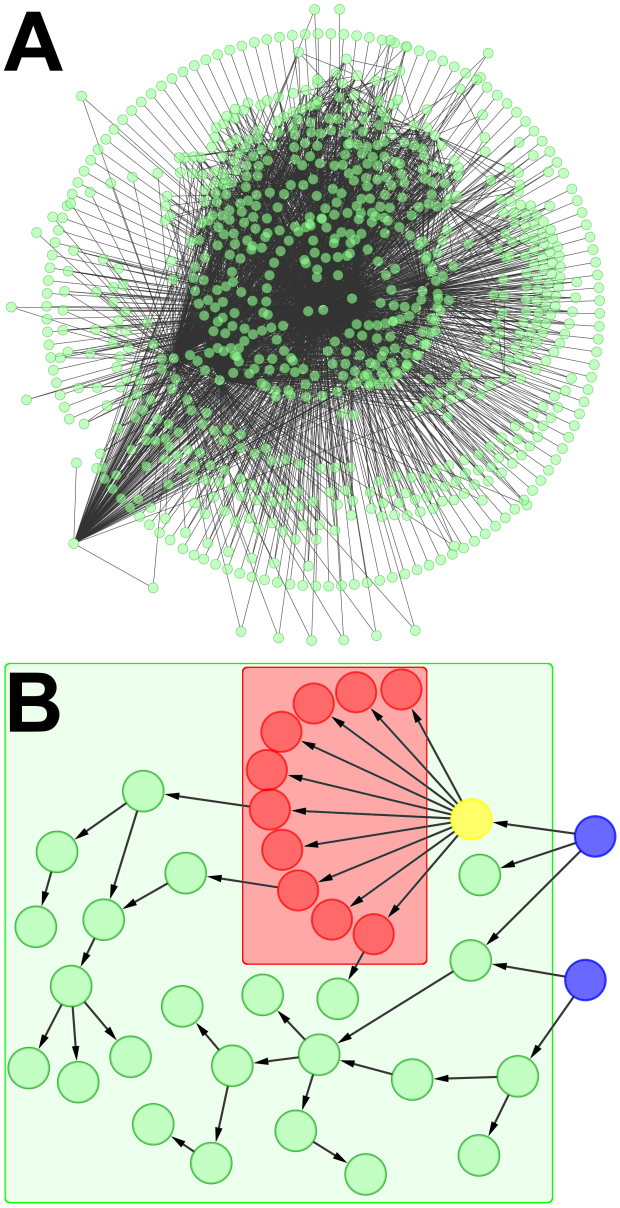

Figure 4.7: **(A)** The expanded M-VBSSM network connections that are in common to both the *B. cinerea* and *P. syringae* infection models. The shared connections comprise roughly 10% of each of the individual networks, suggesting that some regulatory events are shared between the responses but the majority of the observed effects are condition-specific. The network model, whilst very information-dense, is prohibitively large for manual scrutiny, requiring automated functionality analysis. **(B)** Functional inference for expanded M-VBSSM networks. The functionality of a given node of interest (marked in yellow) can be inferred by assessing the overrepresentation of GO terms among its immediate downstream targets (highlighted in red). The procedure can be made more stringent by limiting the context to the analysed network only, setting the universe in the overrepresentation analysis to all the downstream nodes in the network (highlighted in green). The nodes that lack any upstream regulators, marked in blue, are not part of this set.

Table 4.3: The number of TFs in the individual pathogen response networks, as well as the intersection, with predicted downstream functionality

Network	Eligible nodes	Functional nodes	Hormone nodes
B. cinerea	211	132	64
P. syringae	328	173	74
Intersection	45	10	5

To be eligible for functionality analysis, a network node needs to have at least one downstream node, with the total of eligible nodes for each of the models also included in the table in the Eligible nodes column. The Functional nodes column captures the number of nodes that have at least one overrepresented GO term among their downstream targets. Also featured is the number of nodes with GO terms related to at least one of the four primary defence hormones (jasmonic acid, salicylic acid, abscisic acid and ethylene) overrepresented among their immediate downstream targets.

genes, is shown in Table 4.3, while the complete results of the analysis are attached as Supplementary Dataset 9.

In this study we are highlighting TFs predicted to play a role in hormone signalling during the defence response. Hormones play a vital role in the plant's response to a number of external stimuli, including pathogen infection, with crosstalk between the signalling induced by different hormones being vital in fine-tuning the response (Pieterse et al., 2009). *B. cinerea* and *P. syringae* infections are known to elicit greatly different hormonal responses, with jasmonic acid being dominant in the former and salicylic acid in the latter, and the antagonism between the two being well documented (Li et al., 2006; Pieterse et al., 2009). As such, we may detect differences in regulation of hormone signalling pathways between the *B. cinerea* and *P. syringae* networks. Table 4.3 features the number of TFs in the *B. cinerea*, *P. syringae* and intersection networks whose direct targets were enriched for GO terms related to at least one of the four primary defence hormones (jasmonic acid, salicylic acid, abscisic acid, and ethylene). A selection of 12 network nodes, featuring both genes previously annotated with functionality related to the four defence hormones and novel candidates, can be seen in Table 4.4. The functional inference procedure was able to capture TGA3, known to play a key role in jasmonic-acid mediated response to *B. cinerea* infection (Windram et al., 2012), as potentially involved in mediating signalling related to all four hormones in that response. This is possible, due to the high degree of hormone signalling crosstalk in stimulus response (Pieterse

Table 4.4: The twelve selected genes and their prior (GO terms present in annotation) and inferred (result of functional analysis performed in the study) defence hormone related functionality

AGI	Name	B	P	Prior Function	Inferred Function
AT5G50010					JA, SA, ABA, ethylene
AT5G15850	BBX2				JA, SA, ABA, ethylene
AT3G18990	VRN1				JA, SA, ABA, ethylene
AT1G09030	NF-YB4				JA, SA, ABA, ethylene
AT1G22070	TGA3			JA, SA	JA, SA, ABA, ethylene
AT1G25550					JA, SA, ABA, ethylene
AT3G01970	WRKY45				SA, ABA, ethylene
AT4G39780				ethylene	JA, ABA, ethylene
AT1G76110				JA, SA	JA, SA, ABA, ethylene
AT2G20825	ULT2				JA, SA, ABA, ethylene
AT1G80590	WRKY66				JA, SA, ABA
AT2G16720	MYB7			JA, SA, ABA, ethylene	JA, SA, ABA, ethylene

A number of the putative regulators (MYB7, AT1G76110, TGA3, AT4G39780) were previously annotated to have a role in some of the processes the functional analysis inferred, whilst the other five genes appear to be novel. Columns B and P represent *B. cinerea* and *P. syringae* infection respectively, with a green shaded region in the corresponding column highlighting the stimuli where each of the genes exhibited functionality relevant to the analysis. The three genes with both the B and P fields shaded had their listed functionality identified in the intersection network.

et al., 2009). The selection of 12 nodes features 5 exhibiting their functionality in *B. cinerea* infection, 4 exhibiting their functionality in *P. syringae* infection, and 3 exhibiting their functionality in the intersection network.

4.2.5 Identification of a combinatorial regulation network module predicted to play a role in hormone signalling

A small network module has been identified in the intersection network, containing five transcription factors with a predicted role in the response and signalling to all four hormones in both defence responses. A representation of the module, along with the highlighting of the hormones that the downstream genes each individual network node is involved in, can be seen in Figure 4.8. There is a high degree of connectivity between the nodes, which in conjunction with a number of similarities in downstream targets (outlined in Table 4.5) suggests that this network module represents combinatorial regulation of hormone signalling.

An examination of the network module structure, in conjunction with the

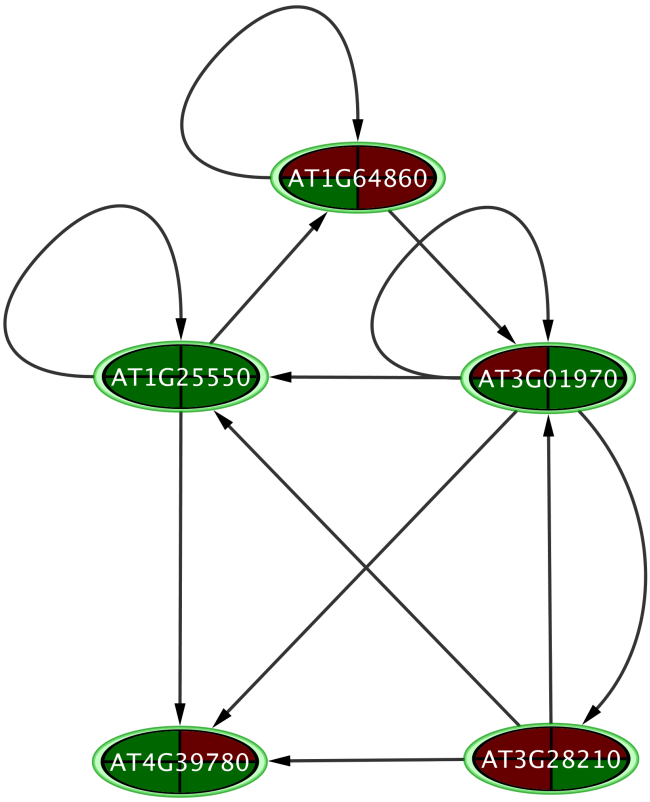

Figure 4.8: A truncated view of the five-gene combinatorial regulation hormone signalling interaction identified in the *B. cinerea*/*P. syringae* intersection network. Each of the nodes is divided into four areas, corresponding to jasmonic acid, salicylic acid, abscisic acid and ethylene starting in the top left area and going clockwise. A green shading of the appropriate area indicates that the node has GO terms related to the relevant defence hormone overrepresented among its immediate downstream targets, which are absent from the visualisation for clarity.

Table 4.5: The number of shared downstream targets between the five genes forming a hormone signalling module in the intersection network

Node	Targets	AT1G01970 306	AT1G64860 318	AT4G39780 9	AT3G28210 129
AT1G25550	24	21	19	1	19
AT3G10970	306		160	6	77
AT1G64860	318			7	120
AT4G39780	9				5

The consistent overlaps in membership, in combination with the high connectivity of the five module genes themselves, suggests combinatorial regulation action in non-JA hormone signalling, with AT3G01970 serving as the hub of the network and the other nodes serving as fine tuners. In the case of JA, the network reduces to a sparse cascade in AT1G25550 and AT4G39780, with both of them having few targets and commonly regulating only one gene.

downstream target overlaps, reveals WRKY45 (AT3G01970) as the main hub in SA, ABA and ethylene signalling. WRKY45 possesses a high number (306) of downstream targets, but manages to retain overrepresentation of GO terms related to three defence hormones among its downstream nodes. AT1G25550, the only node to have all four defence hormones overrepresented among its downstream targets, only regulates 24 genes. However, 21 of those are in common with WRKY45, so it seems likely that AT1G25550 acts as a fine tuner of the hormone signalling that WRKY45 has the potential to induce. WRKY45 also maintains a high degree of overlap (at least 50% of the other node's downstream genes) with the other three nodes present in the network module, indicating those genes' roles in fine tuning appropriate hormone signalling. Of particular note is AT4G39780, which only has 9 downstream targets but manages to retain overrepresentation in both SA and ABA, indicating a similar fine-tuner role to AT1G25550. An examination of the ABA-overrepresented SAP12 (AT3G28210) and the ethylene-overrepresented SIGA (AT1G64860) reveals a surprising trend, where 120 of the 129 SAP12 downstream targets are featured among the SIGA downstream targets, but neither node is overrepresented for the other's hormone functionality. As such, it is likely that the 198 targets that SIGA did not have in common with SAP12 are the main ones with ethylene related functionality and dilute the ABA overrepresentation to the point of it not being detected.

An examination of the network module from a JA point of view reveals an intriguing and entirely reverse trend. JA is the only hormone that WRKY45 does not have overrepresented among its downstream nodes, and it is only present in

AT1G25550 and AT4G39780, with both of those having a very small number of downstream targets. The model becomes even more interesting when it is noted that of the 24 targets of AT1G25550 and 9 targets of AT4G39780, only one is in common. As such, the network module seems to move away from its standard combinatorial signalling with a high number of downstream targets when exposed to JA, with a small scale but very causal model in place. Said small scale model can be repurposed into an WRKY45 signalling fine tuner when other hormone signalling is involved, directly linking the crosstalk in the signalling.

4.3 Discussion

The reconstruction of functionally informative regulatory networks from large scale expression data is a computationally daunting task, often requiring a large degree of compromise in terms of input. Within the article, we have proposed an extension to VBSSM (Beal et al., 2005), an existing network inference algorithm, which makes it possible to mine the full differentially expressed TF space for an underlying regulatory model by subdividing the tasks into smaller scale local models. The method, termed M-VBSSM, was subsequently applied to six time course datasets showing Arabidopsis responding to a number of biotic (*B. cinerea* and *P. syringae* infection) and abiotic (drought, high light, long and short day natural senescence) environmental conditions, inferring the underlying TF-only regulatory interactions. In parallel, the same datasets were mined with Wigwams, a tool for the identification of potentially co-regulated genes across multiple time course datasets, identifying the shared regulatory patterns created by the network crosstalk across two or more of the conditions. These co-regulated downstream target genes, termed network footprints, exhibited a variety of functions across the different modules, including the extension of nucleosome and chloroplast findings made for *P. syringae* data across further stimulus responses (Lewis et al., in press). A number of regulatory predictions were also made for the modules, based on protein binding microarray experimental results (Franco-Zorrilla et al., 2014; Weirauch et al., 2014), revealing the potential regulators driving the identified responses. The Wigwams modules were then used to expand the TF-only regulatory network models, using the co-regulated nature of the genes captured in the algorithm output and the regulatory interactions forming the network model. This significantly expanded the number of nodes and connections in the networks, with the exception being high light with no new nodes or connections added. These expanded network models, prohibitively large to examine by hand, were automatedly individually mined for relevant functionality through

examining the overrepresented GO terms among immediate downstream targets of all nodes captured in the network. This revealed a number of potential genes, both known and novel, possibly involved in primary defence hormone signalling in the *B. cinerea* and *P. syringae* infection response networks, with a follow-up analysis of the intersection of these networks identifying a 5-gene module playing a part in the combinatorial regulation of hormone signalling in the crosstalking network overlap of the two responses.

The choice of a network inference algorithm is always debatable, with the problem being further exacerbated by the scale of the data in the analysis. In terms of raw network reconstruction accuracy, CSI (Klemm, 2008; Penfold and Wild, 2011) was found to perform the best on synthetic data modelled after model organism networks (Penfold and Wild, 2011), but the algorithm scales poorly with the number of genes involved, requiring an indegree run time compromise limiting the evaluated number of parents for each profile, potentially missing the full scope of biological interactions. The proposal of M-VBSSM allowed for the mining of the entirety of the DE TF space through dividing the task into the identification of smaller networks centred on each individual gene, subsequently joining the local models into one final network. The extension of the gene space allowed for the capturing of a complete regulatory skeleton of TF-TF interactions, forming the core of the regulatory signal flow in stimulus response, and these models were subsequently expanded with non-TF genes through Wigwams module integration, resulting in a more detailed model featuring downstream targets that would be far beyond the computational scope of basic network inference algorithms. The application of a similar Metropolis wrapper to CSI could be a possibility, and would be a good avenue for future work, but VBSSM was chosen for this study due to its lack of indegree constraints, explicit capturing of hidden states and reasonable performance in the inference algorithm comparison (Penfold and Wild, 2011). It should also be noted that the resulting network model need not be the definitive underlying set of regulatory interactions, as genes can also be regulated post-transcriptionally in a manner that cannot be reconstructed from transcriptional data. However, given the nature of the data, the reconstruction is as complete as possible.

The footprint of these networks was analysed using Wigwams. The main strength of Wigwams lies in its ability to differentiate between statistically significant, dependent co-expression indicative of co-regulation, and independent co-expression merely stemming from the relative abundance of particular profiles. As Wigwams mines multiple time course datasets for co-expressed gene modules, it is able to assess whether the genes appear co-expressed more often than expected by

chance, thus differentiating between the two co-expression types. The Wigwams analysis of the time course datasets reveals a small number of modules spanning drought and high light, suggesting that those are the most mechanically independent of the analysed conditions. This is further consistent with Wigwams modules being unable to augment the high light network with any extra connections. The conditions most consistently showing up in modules together are long and short day senescence, while the highest number of modules in any single condition combination are pairwise *B. cinerea* and *P. syringae* infection modules at 9.

The design of Wigwams focuses on obtaining modules with a high chance of co-regulation, and a natural follow-up is to attempt an identification of the transcription factors that are likely to be responsible for the observed co-regulated behaviour across multiple conditions. A standard approach to identifying such genes involves the identification of overrepresented transcription factor binding motifs in the promoters of gene groups of interest, with this particular application allowing for an expression data follow up and identification of prime potential regulators for experimental validation. Whilst recent advancements to the state of TF binding site knowledge have been made through the application of protein binding microarrays, both in small-scale studies aiming to identify a number of highly accurate motifs for a select group of transcription factors (Franco-Zorrilla et al., 2014) and larger scale work across multiple organisms, supplementing protein binding microarray (PBM) experiments for representative TFs with DNA binding domain (DBD) modelling (Weirauch et al., 2014), our knowledge remains imperfect. This is well evidenced by only 24 of the 71 Wigwams modules having a known transcription factor binding site overrepresented in the promoters of the module members, and cannot be blamed on FDR stringency alone as applying the Benjamini-Hochberg correction instead of Bonferroni still results in only 37 modules having at least one significantly overrepresented transcription factor binding motif in their members' promoters. Our knowledge of TF binding sites is going to continue improving, with the development and application of PBMs being a major stepping-stone in the process.

The performed network expansion procedure resulted in the inclusion of key downstream genes into the network models, adding functional information to the TF-TF interaction foundation. The use of Wigwams to produce the gene modules used in the expansion procedure could be debated, due to Wigwams modules spanning two or more conditions whilst the TF-only models were inferred for each stress independently. As such, the functionality added into the network is limited to the footprint of the crosstalk between networks for multiple conditions. An intuitive solution to the problem would be the application of an algorithm capable of mining

a single dataset for groups of co-regulated genes, but no existing method is capable of the degree of discrimination between dependent co-expression indicative of co-regulation and independent co-expression merely resulting from the abundance of profiles that Wigwams is. The algorithm that comes closest is CCC-Biclustering (Madeira et al., 2010), but the heavy discretisation of the expression profiles and localisation of the statistical evaluation make the resulting modules less informative than those identified by Wigwams.

Another issue with the network expansion procedure include the adjustment of the co-regulation overlap, resulting in the addition of a Wigwams module to the network model. The required overlap between a network node's downstream targets and a Wigwams module's gene members was set to four based on testing on the *B. cinerea* and *P. syringae* infection networks. This threshold resulted in the new connections roughly doubling the size of the network for those two models. In terms of actual biological models, this is probably an insufficient number of downstream genes — the majority, if not all, of differentially expressed genes are driven by some elements of the identified network, yet in the case of *B. cinerea*, only 2623 of the 9838 differentially expressed genes are present in the expanded network model (Windram et al., 2012). Once again, some of this observed downstream gene limitation can be blamed on the crosstalk aspect of Wigwams modules, but in addition to that a conscious attempt was made to preserve the underlying regulatory structure of the original TF-TF models in the expanded networks and avoid overly flooding them with new downstream target links. Applying the same parameters to the expansion of other networks yielded diverging results, with long and short day senescence gaining a high number of new connections, resulting in models likely to be closer to the actual biological flow of information, whilst high light wasn't expanded with any nodes or connections. This may stem from high light being a relatively unique response, showing up quite scarcely across Wigwams modules, and the inferred TF-TF network potentially focusing more on an individual response instead of the elements of crosstalk whose footprints were detected by Wigwams modules.

In the case of network nodes with no downstream targets, the network expansion procedure has no way of adding downstream targets due to the lack of co-regulation overlap with a Wigwams module. Attempts to combat this issue could include running network inference with profiles representative of downstream target behaviour, such as Wigwams modules, present, but this would increase computational load. In the case of Wigwams modules, this would bring up the question of optimal module selection to compromise computational tractability, and the application of other algorithms would run into the same regulatory assessment problems

that have been previously discussed.

Whilst the focus of the analysis in the paper are the *B. cinerea* and *P. syringae* intersection networks, the presence of the four other datasets was beneficial and allowed for a more informative expanded network to be constructed. This stems from the inherently multi-condition nature of the modules Wigwams identifies, and by including the additional datasets in the analysis it was possible to identify modules spanning one of the defence responses and some other, unrelated dataset, with those modules being subsequently used for downstream target integration into the individual defence response networks. In the interest of computational efficiency, if only a select few of the available time course datasets are of interest from a regulatory network perspective, TF-only network inference can be limited to them whilst mining more datasets with Wigwams to capture modules in a similar manner. All of the expanded networks, not merely the defence response ones, are featured in Supplementary Dataset 8 and can be freely examined for relevant connections or functionality.

The performed analyses are very rich in immediately experimentally testable biological hypotheses — a number of Wigwams modules have putative regulators predicted based on TF binding motif overrepresentation followed up with timing of differential expression inferred from experimental data, the expanded network models have functional information identified for nodes with downstream targets, and the five gene network module performing combinatorial regulation of hormone signalling in the intersection network is a succinct set of interactions potentially fine-tuning downstream regulation. An informative approach would be the use of knockout lines to assess the predicted functionality. However, the predictions may fall prey to the main hurdle in the transition from genotype to phenotype — robustness and redundancy in signalling (Xu et al., 2006). As evidenced by the network module, it is possible for a number of TFs to share downstream targets, and in the event of one of them losing the ability to carry out its regulatory role, one of the others would be able to compensate for its absence. It seems likely that knocking out WRKY45 would not have an immediate effect on hormone signalling functionality, as the other genes in the module would likely be able to carry out its signalling roles through the overlap in the downstream genes. However, a line with multiple genes from the interaction knocked out may produce more relevant results, with a number of the redundant components removed and the regulation potentially becoming disrupted.

4.4 Methodology

4.4.1 Inference of TF-only Network Models

Genes encoding known or potential transcription factors were identified within the genes differentially expressed in each time series data set and are given in Supplementary Dataset 1.

M-VBSSM was performed for each of the time course datasets (Arabidopsis response to *B. cinerea* and *P. syringae* infection, drought, high light, long day and short day natural senescence), with the input being all of the transcription factors differentially expressed in that dataset. Let G be the set of transcription factors that the network inference is being performed for, and X be the single transcription factor that this particular instance of M-VBSSM modelling is centred on. Let J be the set of genes the current VBSSM model includes with complementary set $J' = G \setminus J$, with $L(J)$ being the corresponding model's likelihood. A new set of genes J_{new} can be created by switching N random genes from J with N random genes from J', where $N \sim \mathrm{B}(\min(|J'|, |J \setminus X|), p)$ with p being a random variable that can be tuned for optimal acceptance rates. The new model with the J_{new} gene set can then be compared to the prior model for acceptance or rejection using the Metropolis rule:

$$p(accept) = \min\left(1, \frac{L(J_{new})}{L(J)} \frac{p(J|J_{new})}{p(J_{new}|J)}\right)$$

As the genes to be swapped are chosen at random from their sets, they represent random samples from hypergeometric distributions. The transition probabilities are products of binomial and hypergeometric distributions and cancel out.

Every gene in a given dataset was used to construct a local network model according to the procedure described above with $|J| = 80$. 2000 iterations of Metropolis sampling was performed for each of the genes, using the VBSSM marginal likelihood to assess the quality of the fits, and the final local model had its $CB + D$ parameter matrix converted to Z-scores, to which a significance threshold of 1.65 was applied to determine a local network. Once this procedure was performed for all of the genes, a union of all the local networks was performed to obtain a final M-VBSSM network for the dataset. This procedure was performed for all six of the time course datasets.

4.4.2 Identifying Wigwams modules

The six time course datasets were mined for Wigwams modules (Polanski et al., 2014), keeping all the parameters as specified in Polanski et al. The initial mining procedure returned 10586 modules spanning two to six conditions, with an average module size of 22 genes. This resulted in 71 final Wigwams modules after removing redundancy, whilst only reducing the number of unique genes found within the modules from 5925 to 5434, with an average module size of 168 genes.

4.4.3 Functional inference of Wigwams modules

The 71 modules were mined for biological information by assessing full GO term overrepresentation with BiNGO (Maere et al., 2005) and known transcription factor binding site overrepresentation. Arabidopsis transcription factor binding sites were obtained from protein binding microarray experiments conducted by Franco-Zorrilla et al. (2014) and Weirauch et al. (2014), and their occurrences in 500 bp Arabidopsis promoters were identified through FIMO (Grant et al., 2011) ran with default parameters. In the case of both GO term and binding motif overepresentation, the hypergeometric test was used to assess statistical significance of the overrepresentation with the significance threshold of 0.05. For the GO term test, the Benjamini-Hochberg FDR correction was used, whilst the Bonferroni correction was used for the transcription factor binding site overrepresentation analysis. In order to validate putative regulators, the Gaussian process gradient tool (Breeze et al., 2011) was used. For modules, the first time point at which at least 50% of the module's members exhibited a change in expression identified by the Gaussian process gradient was defined as the single change point and used to compare the module's expression trend against that of the putative regulator.

4.4.4 Network expansion and functional analysis

For the purpose of network expansion, the initial 10586 modules were used, as these are the smallest, and most stringent modules. If a Wigwams module contained 4 TFs predicted to be co-regulated by TF-X in the M-VBSSM network, then the remaining non-TF members of the module were added to the network as additional targets of TF-X. In additional, if a single TF was predicted to be regulated by TF-Y in multiple stress conditions, and a Wigwams module containing the single TF spanned the same stress conditions, then the other members of that module were added to the network as targets of TF-Y. To place emphasis on the original network structure, the added Wigwams module connections were not themselves

used in the expansion process. Functional inference was performed by testing GO term overrepresentation among genes immediately downstream of the given node of interest, with Benjamini-Hochberg used for FDR correction (adjusted $p<0.05$). In order to maximise stringency, the GO term gene universe was limited to all the downstream nodes of the given analysed network.

Chapter 5

Discussion

5.1 Significance of the work

The featured work, encompassing the development of Wigwams and its subsequent joint application with other methods to sets of time course datasets, allowed for the formulation of numerous biological conclusions. The design of Wigwams lends itself to a higher degree of confidence when describing detected co-expression as co-regulation. The presence of both the virulent and avirulent *Pseudomonas syringae* pv. *tomato* DC3000 strains in the experiment led to novel findings on the mechanics of effector-driven plant defence response suppression. The enhancement of transcription factor-only regulatory networks lends itself to easy formulation of precise functional hypotheses for experimental validation.

Whilst pre-existing methods allow for the identification of genes exhibiting co-expression across multiple time course datasets (Heard et al., 2005; Supper et al., 2007), each of the pre-existing algorithms features its own weaknesses, and on top of that none of these methods attempt to assess whether the co-expression trends they detect are occurring by chance, merely stemming from a high abundance of the detected profiles in each of the individual datasets and not carrying any regulatory implications, or possessing a shared regulatory mechanism across the conditions. Wigwams utilises the fact it mines multiple time course datasets by statistically assessing whether the detected co-expression is occurring more often than would be expected by chance, and uses this information to discriminate between dependent co-expression indicative of co-regulation and independent co-expression merely stemming from profile abundance. This is novel for multiple time course dataset analysis; CCC-Biclustering evaluates the probability of the discretised expression profiles of the modules it detects occurring by chance as well, but this method is

only capable of analysing a single time course dataset (Madeira et al., 2010).

Wigwams can be applied to a number of different tasks, with a number of both standard and atypical applications shown in Chapters 3 and 4. When applied to *P. syringae* infection data, Wigwams was able to identify a consistent down-regulation of chloroplast-related genes shared across the response to both the virulent and avirulent strain, likely as part of PTI, and its joint application with transcription factor binding motif analysis elucidated the potential of combinatorial TF action mediating abscisic acid response in a manner parallel to rice (Hobo et al., 1999). When applied to a higher number of high resolution time course datasets showcasing *A. thaliana* response to a number of biotic and abiotic stimuli, Wigwams was able to extend the *P. syringae* infection findings of nucleosome and chloroplast down-regulation to other conditions, and when combined with single transcription factor binding motifs was able to establish WRKY33 as a key regulator of a module exhibiting co-regulation across *Botrytis cinerea* and *P. syringae* infection, further ascertaining its role as a key component of the regulation of the jasmonic acid and salicylic acid response (Birkenbihl et al., 2012).

An additional application of Wigwams was its use to enhance M-VBSSM transcription factor only regulatory models due to a high degree of certainty that the identified modules are co-regulated. This in turn allowed for network node functional inference in a novel manner and the subsequent identification of a putative five-gene interaction shared across *B. cinerea* and *P. syringae* infection. This small-scale regulatory interaction could play a key role in the fine-tuning of signalling of all four primary defence hormones across the two responses.

The *P. syringae* analyses not involving Wigwams were also able to answer a number of biological queries. Applying the gradient tool to the differences of expression profiles allowed for the functional assessment of the timing of the response, and utilising the single time point differential expression information led to the identification of early sustained differential expression profiles and the detection of effector-driven down-regulation of nucleosome genes in the virulent strain. The network model inferred jointly across all three of the time course datasets produced a core set of interactions fine-tuned in a condition-dependent manner, with the subsequent analysis of hubness (assessment of the number of downstream targets each network node is predicted to regulate) identifying a number of putative genes playing a key role in the effector-specific differential response to the virulent and avirulent strains.

5.2 Discussion of the applied methodology

In the case of the Wigwams analyses, one of the key factors influencing the final outcome is the choice of the distance metric used to compute the similarity or lack thereof between pairs of gene expression profiles. The results are captured in a massive three-dimensional correlation matrix, which is subsequently mined for instances of statistically significant co-expression in a complete per-gene search. The choice of metric to compare time series expression profile similarity is a known issue in time series analysis, and a series of comprehensive testing was performed on a number of popular distance metrics across a set of different time course experiments to assess the frequency with which individual metrics were capable of identifying gene pairs with an underlying experimentally verified relationship (Yona et al., 2006). The testing revealed the authors' novel mass distance metric, which accounts for the positioning of the expression values of the tested gene pair in the overall expression value distribution at each individual time point, nearly uniformly outperforming all other metrics across the examined time course datasets. However, the Pearson Correlation Coefficient (PCC) was chosen for the Wigwams analysis due to its greater computational tractability and ease of result interpretation, in conjunction with a stable performance in the metric testing. Early attempts were also made with the Spearman Correlation Coefficient, but the resulting modules were of lesser quality than those identified with the aid of PCC, making the findings partially consistent with the metric testing where the Spearman Correlation Coefficient performed extremely erratically. Attempts to branch away from PCC saw the application of another one of the authors' novel metrics, EucPear (Yona et al., 2006), to an experimental small-scale Wigwams analysis not featured in any of the prior chapters where the actual level of expression of each gene was of key importance. Whilst performing well for the task at hand, the significant computational increase due to averaging multiple computations of the non-deterministic metric to get around the significant output spread makes it unfitting for the current Wigwams implementation. If computational tractability were of no concern, the mass distance metric would merit testing to assess its impact on the results.

Whilst the tractability of Wigwams was a factor in the process of choosing a distance metric, with the final version capable of mining a number of time course datasets for modules in a matter of hours, computational tractability played a far greater role in the development and application of network inference algorithms to the data. The modelling captured in Chapters 3 and 4 demonstrates an evolution of the tractability and availability of network inference across the duration of the

project. As evidenced by the analysis performed on DREAM4 data, allowing for the comparison of the reconstructed network models against the known regulatory interactions that were used to produce the test data using stochastic differential equations, CSI is the most accurate (Penfold and Wild, 2011). However, at the time the implementation of CSI was severely limited in terms of the number of profiles it could analyse, and whilst it was possible to utilise the algorithm to infer a network model for a set of profiles for 44 representative clusters and a pathogen growth curve (Windram et al., 2012), application to large-stale network reconstruction was out of the question. At the time, none of the network inference algorithms were capable of handling a problem of such scope, which subsequently led Dr Christopher Penfold to the creation of the Metropolis wrapper for VBSSM as discussed in Chapter 4. By focusing on one gene at a time and swapping a binomially distributed number of random genes between each iteration, it was possible to break up the large-scale problem into more manageable pieces and produce a model by amalgamating the local results. Since then, the Matlab implementation of CSI was improved, greatly reducing run time and making it possible to analyse problems of scope far extending what the old network inference algorithm implementations could handle (Penfold et al., 2015). This improved implementation was utilised to infer the joint three-condition network model shown in Chapter 3, with the optimised code quickly handling 619 input profiles (618 early response transcription factors and a splined pathogen growth curve) with three replicates on a dedicated computing cluster. The M-VBSSM network models for Chapter 4 were kept, as the underlying methodology is scientifically sound and the Wigwams expansion analysis ended up producing a number of interesting results. Nevertheless, redoing the analysis utilising CSI models inferred for each of the datasets would be an interesting future avenue and could yield other relevant conclusions not covered by the current analysis. In terms of further tractability improvements, it should be noted that all of the tools are still only currently available as Matlab implementations. Matlab, whilst being a very comfortable environment to program in, is nowhere near as computationally efficient as lower level computational languages with regards to most operations. As such, the efficiency of CSI and other network inference algorithms would further increase if they were to be recoded into C or even Python, leading to the possibility of slightly relaxing the indegree limitation imposed for computational tractability whilst retaining acceptable performance time, in turn allowing the direct capturing of more complex combinatorial regulatory interactions. The recoding could also be expanded to Wigwams, with the greatest computational bottleneck being the creation of the correlation matrix, but this is a less pressing issue due to

the current Matlab implementation being reasonably efficient. Nevertheless, moving the algorithm to a lower level programming language could accommodate the aforementioned mass distance metric testing, likely with a slight hit to user flexibility as any desired custom distance metrics would have to be provided in whatever language Wigwams would be recoded to instead of Matlab.

In terms of method selection, the most debatable one is the application of single time point differential expression information to the identification of sustained early effector-driven expression change in the *P. syringae* infection response analysis. Whilst a valid approach in its own way, and yielding interesting conclusions with regards to nucleosome down-regulation and the complementary lack of up-regulated functionality potentially stemming from the resulting transcriptional destabilisation, applying a single sample algorithm to each time point individually fails to utilise the temporal nature of the data to its fullest potential. It would be preferable to conduct this analysis with a proper time course algorithm, and attempts have been made with the time-local version of GP2S (Stegle et al., 2010). However, the obtained results proved to be unsatisfactory, predominantly due to the sharp wound response present at the start of the time course. The initial drastic shift in transcript levels stems from the syringe inoculation, introducing the bacteria to the plant interior and triggering its endophytic stage while wounding the plant in the process. Whilst the wounding itself is undesirable, the syringe inoculation process is the most reliable way to introduce desired concentrations of the bacteria into the plant, as discussed in Chapter 3. This results in the early time points of the *P. syringae* dataset showing a plant wound response, in addition to any possible host-pathogen interactions that may be taking place. The vast majority of genes exhibit uniform behaviour across the mock and infection with both strains at this wound response stage, but the scale of the expression shift often dwarfs the expression changes across the remainder of the time course and affects the Gaussian process fits in an adverse and usually non-uniform manner, leading to the algorithm proclaiming that the gene is immediately differentially expressed due to slightly different magnitudes of the edge effect for control and infected fits. Accounting for this phenomenon by ignoring the differential expression score up to the first local minimum, theoretically compensating for the edge effect and giving the fit room to stabilise, doesn't fully solve the problem, as the entirety of the fits are warped by this drastic initial change. Removing the first time point leads to information loss, as there are a number of genes exhibiting early differential expression that would now form the edge of the fit. An ideal solution would involve the application of an algorithm accounting for the temporal nature of the data capable of assessing

the differential expression status of the gene at every time point without being as troubled by the early wound response. A promising candidate is DEtime, which operates in a manner similar to GP2S without the time-local model sampling — a perturbation time is defined, with a single Gaussian process fit to the data prior to the perturbation time and two condition-specific Gaussian processes fit to the data upon reaching the perturbation time. The method has shown promising results on a trial run of a sample of the *P. syringae* data (Yang et al., in press). However, the algorithm only identifies the time of first differential expression of the gene without assessing its subsequent expression trends, making it potentially vulnerable to brief bursts of differential expression and rendering the current version of the method unfit to handle identifying sustained differential expression, as was the goal here.

The biological interpretation of the computational inference results is reliant on the current state of knowledge of individual gene functionality and transcription factor binding specificity. For the purpose of making computational functional analyses easier and more uniform to perform, possible gene functionality is stored in a branching graph of functionality specialisation, known as Gene Ontology (GO) terms (Ashburner et al., 2000), with a typical analytical application being scanning gene groups for GO term overrepresentation with the aid of tools such as the Cytoscape plugin BiNGO (Maere et al., 2005). Ideally, assigning a GO term to a gene should be based on dedicated experimental validation, but this becomes an extremely daunting task to carry out even for model organisms when faced with tens of thousands of genes. As such, a number of GO term annotations are based on computational inference. An example is the development of a large-scale model utilising expression data, experimentally validated relationships in the form of protein-protein interactions and targeting by transcription factors, and pre-existing functional information, with the model subsequently divided into modules and per-module functional inference predicting novel roles for over 5000 genes (Heyndrickx and Vandepoele, 2012). These predictions tend to have a decent degree of accuracy, as evidenced by follow-up experimental assessment for AraNet validating two of three novel functional hypotheses (Lee et al., 2010). However, they are still only computationally inferred predictions, and it would be preferable to validate them experimentally to have a higher degree of certainty in the GO term annotations. As mentioned previously, this would be an experimental undertaking of enormous scope, with approximately 40% of *A. thaliana* genes having some aspect of their functionality experimentally assessed, but only an eighth of those possess experimental evidence for their biochemical activity, subcellular location and biological role. Computational inference allows for the approximation of the role of up to 95% of *A. thaliana*

genes (Rhee and Mutwil, 2014).

The second avenue for biological interpretation utilises information on transcription factor binding sites and specificity. If given a group of co-regulated genes, an examination of their promoters with the aim of detecting overrepresented recurring motifs, similar to scanning for overrepresented GO terms, can provide insight into the shared regulatory mechanisms driving the examined gene group, with a higher specificity of knowledge on the identified motifs enabling more precise regulatory hypotheses. Analyses performed over the course of the project showcase an advancement of the state of knowledge on transcription factor binding motifs — Chapter 2 utilises a collection of 349 motifs found in PLACE (Higo et al., 1998) and TRANSFAC (Wingender et al., 1996) databases. Both of these databases store motifs identified in published scientific work, but the individual studies are independent and the methodology of identifying motifs is heterogeneous. Additionally, the motifs span a variety of organisms, and the most reliable way to translate the findings to *A. thaliana* is to base on transcription factor families. Whilst this provides some degree of specificity to the motifs, a more precise prediction is preferred. During the course of the project, a number of dedicated protein-binding microarray (PBM) studies were conducted by other research groups to elucidate the specificity of binding of individual transcription factors. A smaller scale study identified up to three motifs for each of the 63 analysed *A. thaliana* transcription factors (Franco-Zorrilla et al., 2014), whilst a larger scale experiment identified motifs for over 1,000 transcription factors from multiple organisms, with 222 of the inferred motifs stemming from *A. thaliana* (Weirauch et al., 2014). These PBM-derived single transcription factor binding motifs were utilised in the analysis of Wigwams modules in Chapter 4, and the higher specificity allowed for the formulation of highly focused regulatory predictions when combined with a local follow-up analysis, such as the identification of WRKY33 as a putative regulator of a group of genes co-regulated across *B. cinerea* and *P. syringae* infection, fitting well with its known role as a key element in the regulation of jasmonic acid and salicylic acid response (Birkenbihl et al., 2012). The previous state of knowledge on transcription factor binding would not have made a prediction that precise immediately possible. The new methodology makes it easier to focus on individual transcription factors in binding site identification — given the fact that there are 74 WRKYs in *A. thaliana* (Pruneda-Paz et al., 2014), singling out WRKY33 as the potential regulator would have been considerably more difficult using just a consensus motif for the whole WRKY family. In spite of these advancements, our knowledge of transcription factor binding is still lacking, as evidenced by none of the predicted regulators from three different transcription factor

families for the *P. syringae* infection and long day senescence module with defence hormone functionality passing the follow-up analysis. Attempts to get around the limited state of knowledge can be made by utilising tools like MEME-LaB (Brown et al., 2013), which mine promoters for overrepresented novel motifs. The results of this type of analysis are the opposite of those offered by scanning for known PBM sequences — all of the potential motifs overrepresented in promoters are detected, but no information is available as to the specificity of transcription factors binding to them. An alternate approach to PBMs in terms of experimental derivation of regulation carried out by a transcription factor can be carried out with ChIP-seq, which can identify where a given transcription factor binds on a genome-wide scale. Whilst providing direct experimental insight into binding targets, ChIP-seq information is scarcer than PBM motifs for single transcription factors — in 2014, there was publically available ChIP-seq data for 27 transcription factors in *A. thaliana* (Heyndrickx et al., 2014). The scarcity of such data can be attributed to a number of factors, with a major issue being the difficulty and rigorous testing needed to obtain the antibodies (Park, 2009). It should be noted that all such cis-regulatory element analyses are greatly dependent on defining the promoter region of interest, with different definitions leading to the identification of binding motifs present different distances from the transcription start site. For the work carried out here, unless otherwise noted, the promoter was defined as the 500 base pairs upstream of the transcription start site. Other studies have used different promoter lengths, such as 200 base pairs (Zhang et al., 2006) or 1000 base pairs (Maruyama et al., 2004).

5.3 Future work

It should be noted that whilst the current implementation of Wigwams features a number of improvements over the original version, there are still areas of the algorithm that could be fine-tuned to yield optimal output, with the definition of optimality varying based on the task at hand. The number of unique genes identified as part of modules saw a drastic increase in the new version of the algorithm — the data Wigwams was applied to in Chapter 4 was previously mined with the original implementation of the algorithm, yielding 3,397 unique genes in its modules (Rhodes, 2012), whilst the current version features 5,434 unique genes in the modules it identifies, an almost 60% increase. However, the original implementation never merged the modules it identified, instead performing an operation called pruning where a dominant module representing a given regulatory phenomenon, determined

based on p-value, would remove redundant modules. This led to a loss of information as mirrored in the final unique gene count, but would produce smaller modules limited in size by the set list used for the hypergeometric test. Merging, the current approach to handling redundancy among modules spanning the same condition span, does a superior job of preserving unique genes but comes at the cost of some of the final modules reaching very large gene totals due to shared widespread regulatory phenomena between conditions. The largest module identified in the Chapter 4 analysis features 1645 genes potentially co-regulated across long day and short day senescence. Based on the nature of the task at hand, it may be preferable to have smaller modules to work with, and then the smaller-scale, representative modules produced by pruning may be of more use. As such, an implementation of pruning is included as an option in the Wigwams GUI. Nevertheless, due to pruning's tendency to lose a high number of genes from the unique gene count, a potential avenue for further Wigwams development would include the proposal of a different redundancy removal procedure, or fine-tuning the parameters of the current approach, to produce smaller modules than merging without the information loss of pruning, preferably without fracturing the statistically significant co-expression identified in the initial mining step.

The second step in redundancy removal, novel to the current implementation of Wigwams, is the handling of modules spanning subsets of conditions but featuring similar genes (termed sweeping). This was not an issue in the original Wigwams implementation, as the statistical framework was limited to testing condition pairs and would only reconstruct modules spanning three or more conditions based on the membership of corresponding pairwise modules. Removing redundant modules spanning different conditions is a difficult task that tries to balance a focused, cohesive output with information loss. Examining modules 13, 26, 28, 44, 50 and 54 of the Wigwams analysis in Chapter 4 shows the perfect example of the current implementation of sweeping in action. Module 54 is a 258-gene module spanning four conditions (*B. cinerea* infection, *P. syringae* infection, long day and short day senescence), and sweeping used it as the base to remove any modules spanning a subset of these conditions whose membership is made up in at least 50% of genes within the four-condition module. The only three-condition module that does not get removed in this step spans *B. cinerea* infection, long day and short day senescence (module 44, 672 genes), and performing sweeping with this module as the base removed all the corresponding pairwise modules apart from the long day and short day senescence one (module 13, 1645 genes). In turn, the absence of the other three-condition modules results in the other pair of conditions modules (modules 26,

28 and 50, 640 to 860 genes) surviving sweeping easily as the four-condition module is unable to remove them. The complex downstream module removal, as outlined above, is a strength of the current implementation of sweeping due to its ability to assess the redundancy of the module structure, but in spite of the procedure there are still a high number of genes featured across all six of the modules mentioned above. This creates an opening for the fine-tuning of sweeping to produce concise output where genes are featured as co-regulated across as high a condition span as possible. The simplest way to achieve this effect would be to remove the gene overlap between a module spanning more conditions and a module spanning fewer conditions (for example, between the four-condition module 54 and the three-condition module 44) from the module spanning fewer conditions. However, this would disrupt the membership of the individual statistically significant co-expression events that were subsequently merged into larger modules, so it could be viewed as not desirable. Another possible approach would be the application of a highly tuned form of sweeping before the initial module merging procedure.

It should be noted that time course expression datasets can feature some degree of structure in the individual dataset conditions spanned. The most prominent example would be AtGenExpress, a set of short *A. thaliana* abiotic condition time course responses with the experiments replicated for root and shoot tissue (Kilian et al., 2007). The current implementation of Wigwams is incapable of applying any degree of classification to the datasets it accepts on input, and the example modules generated by applying Wigwams to a subset of the AtGenExpress data in conjunction with *B. cinerea* and long day senescence time course datasets in Chapter 2 did not attempt to account for the root or shoot nature of the individual time courses. This classification can be easily handled by tensor methods (Li et al., 2012; Zhang et al., 2012), with those approaches capable of dealing with problems of even higher dimensionality (more detailed classification of the individual time courses) in their mining. These approaches suffer from similar drawbacks to typical biclustering tools, and are incapable of assessing whether the co-expression they detect is indicative of co-regulation. Nevertheless, the design space pioneered by these approaches, in conjunction with the existence of datasets such as AtGenExpress, highlight an interesting putative avenue of Wigwams expansion. It would be interesting to see classification of the time courses accounted for directly in the statistical framework, with some possible example AtGenExpress module types including co-regulation across multiple conditions in both root and shoot, and root or shoot individually.

The inferred network models, currently utilised for insight on the interactions and key regulators through the analysis of hubness in Chapter 3 and overrepresented

functionality among downstream nodes in Chapter 4, could be mined for further information. At the moment, the only direct integration of network topology into the analysis is the aforementioned hub assessment, whilst the degree of interconnectivity of the network can carry a lot of information with regards to the nature of its action when triggered by the stimulus (Windram et al., 2014). The current network model in Chapter 3 is quite sparse, and in contrast the individual models from Chapter 4 are highly interconnected, but this stems from differing stringency thresholds due to the vastly different roles the models play in both of those analyses — Chapter 3 sees the network analysed manually for any interesting interactions and has a high stringency threshold, whilst Chapter 4 is more lenient in terms of edge selection and subsequently utilises Wigwams modules to further enhance the models and then computationally mine the massive structures for nodes of interest. A dedicated network topology analysis could be conducted, potentially varying the stringency thresholds and observing how the topology and interactions between nodes change, providing insight into the dynamics of the response — for example, examining the topology of PTI (triggered by flg22) and ETI (triggered by AvrRpt2) specific regulatory interactions in *P. syringae* infection response revealed that while there is a high degree of overlap in the individual responses, the application of the common network is vastly different, with PTI utilising synergy between the sectors to amplify the response while ETI is more compensatory, focusing on triggering the response in the face of possible perturbations (Tsuda et al., 2009). As such, analysing the topology of computationally inferred networks can lead to biological insight.

It should be noted that the entirety of the analysis presented here is computational inference, and no attempts have been made to experimentally validate the findings. This is of critical importance if true biological conclusions are to be drawn from these analyses, as the role of computational tools is not providing final results, but the identification and prioritising of hypotheses for biological validation through fitting follow-up experiments. The first possible experimental approach that can be utilised in such follow-up analyses is the assessment of downstream expression and phenotype of mutant lines, i.e. with a particular gene of interest knocked out or overexpressed. This can provide insight into the functionality of the altered gene — upon computationally identifying TGA3 as a putative key regulator in the *B. cinerea* infection response, the exposure of T-DNA insertion knockout lines of the gene to the pathogen resulted in higher susceptibility, validating the inferred functional predictions (Windram et al., 2012). If T-DNA insertion knockout lines are the approach of choice, then it is preferable to perform the experiments on numerous

independent insertion lines knocking out the same gene, due to the possibility of multiple T-DNA insertions occurring per line (Alonso et al., 2003), and by repeating the experiment on multiple independent knockout lines the impact of possible additional insertions is lessened. However, mutant line screening is not guaranteed to provide a definite negation of a particular gene's role in the stimulus response in the case of a lack of phenotype. The underlying regulatory networks are wired for robustness in the face of both internal and external perturbations (Tsuda et al., 2009), so it is possible that if the gene of interest is unavailable a compensatory set of interactions is activated to mediate the desired signal through alternate means. When scanning *B. cinerea* infection and long day senescence networks for nodes whose knockouts showed alterations to the phenotype, the nodes with the highest downstream connectivity often did not result in altered susceptibility when knocked out. The less robust functionality was carried out by 'middle manager' hubs, with fewer downstream targets — genes in this category showed a higher rate of altered phenotype when knocked out (Penfold et al., in preparation). On similar grounds, comparisons of genes differentially expressed in mutant lines with predicted downstream network targets may not be informative. In addition to that, compensatory signalling concerns aside, mutant line differential gene expression analysis ends up capturing a global phenomenon stemming from the signalling cascading through the entire network, and not even early time point capturing around the moment when the gene of interest is predicted to take effect is guaranteed to identify immediate downstream targets.

Validation of immediate downstream targets should be carried out through a variety of experimental approaches. Yeast one-hybrid (Y1H) can be utilised if a single target gene is of interest and the potential binding of multiple upstream regulators is to be assessed in a relatively high throughput manner. In order to streamline the process, a number of libraries featuring yeast strains transformed with a large assortment of transcription factors of interest exist, with a recent advancement being the construction of a genome-scale sequence-verified resource spanning 1956 *A. thaliana* transcription factors (Pruneda-Paz et al., 2014). However, Y1H has its limitations — the experiment is conducted in a synthetic yeast environment with a transcriptional activation domain fused to the transcription factor of interest, not offering a faithful recreation of the cellular environment conditions present in the plant. Additionally, regulatory events can be quite complex, and it is possible that particular instances of binding promoters are only induced when the plant is subjected to a given condition through combinatorial transcription factor action, such as the joint role of VP1 and TRAB1 in mediating abscisic acid signalling in rice

(Hobo et al., 1999). If the assessment of a potential regulator of interest binding to multiple downstream targets is in order, ChIP-seq is the experimental approach of choice (Park, 2009). Additionally, ChiP-seq gets around the limitations of Y1H due to being performed *in planta*, and can be conducted on tissue from plants subjected to the stimulus of interest, helping validate condition-specific regulatory events (Ricardi et al., 2014).

Finally, it should be noted that all of the conducted analyses run into the severe limitation of merely focusing on mRNA levels. Whilst this is very convenient from an experiment design standpoint due to the ease and reliability of quantifying mRNA levels (Schulze and Downward, 2001; Wang et al., 2009), it fails to account for all the possible regulatory interactions happening at later stages. As a practical example, EIN3 is constitutively synthesised and proteasomally degraded under lack of ethylene stimulus, maintaining an mRNA presence whilst being functionally inactive (Guo and Ecker, 2003). As such, the conducted analyses should be treated as basic and exploratory, with any potential sub-networks of interest demanding highly localised, in-depth experimental follow-up extending beyond mRNA quantification. Experimental techniques that can be beneficial in elucidating high-precision regulatory interactions include yeast two-hybrid (Y2H) for identifying proteins capable of interacting with each other (Causier and Davies, 2002) and mass spectrometry for quantifying the levels of proteins present in the cell (Bantscheff et al., 2007). In addition to protein levels and interactions, differential splicing of transcripts should be accounted for, due to the experimentally validated role of alternative splicing in stress response (Palusa et al., 2007). Upon obtaining a sufficient depth of knowledge about the interactions of all components involved in the highly localised regulatory network, a high-precision mathematical model utilising ODEs can be formulated and the conclusions drawn from it can be used to further refine experimental validation, with the perfect example being the novel role of TOC1 inferred from the highly tuned ODE model of the *A. thaliana* circadian clock (Huang et al., 2012).

Bibliography

Aarts, N., Metz, M., Holub, E., Staskawicz, B. J., Daniels, M. J., and Parker, J. E. Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene-mediated signaling pathways in Arabidopsis. *Proceedings of the National Academy of Sciences*, 95(17):10306–10311, 1998.

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell*, 15(1):63–78, 2003.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., and Galle, R. F. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

Adie, B. A. T., Pérez-Pérez, J., Pérez-Pérez, M. M., Godoy, M., Sánchez-Serrano, J.-J., Schmelz, E. A., and Solano, R. ABA is an essential signal for plant resistance to pathogens affecting JA biosynthesis and the activation of defenses in Arabidopsis. *Plant Cell*, 19(5):1665–1681, 2007.

Aktar, W., Sengupta, D., and Chowdhury, A. Impact of pesticides use in agriculture: their benefits and hazards. *Interdisciplinary Toxicology*, 2(1):1–12, 2009.

Allemeersch, J., Durinck, S., Vanderhaeghen, R., Alard, P., Maes, R., Seeuws, K., Bogaert, T., Coddens, K., Deschouwer, K., Van Hummelen, P., et al. Benchmarking the CATMA microarray. a novel tool for Arabidopsis transcriptome analysis. *Plant Physiology*, 137(2):588–601, 2005.

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., and Cheuk, R. Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301(5633):653–657, 2003.

Altman, R. B. and Raychaudhuri, S. Whole-Genome Expression Analysis: Challenges beyond Clustering. *Current Opinion in Structural Biology*, 11(3):340–347, 2001.

Alvarez-Fernandez, R., Penfold, C., Chernukhin, I., Jenkins, D., Bowden, L., Davey, P. A., Matthews, J. S., Ott, S., Denby, K., Wild, D., Rand, D., Beynon, J., Buchanan-Wollaston, V., Baker, N. R., Lawson, T., Bechtold, U., and Mullineaux, P. M. BBX32 controls a network of high light-induced transcription regulators, photosynthesis and dynamic acclimation in mature Arabidopsis leaves. in preparation.

Anderson, J. P., Badruzsaufari, E., Schenk, P. M., Manners, J. M., Desmond, O. J., Ehlert, C., Maclean, D. J., Ebert, P. R., and Kazan, K. Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *Plant Cell*, 16(12): 3460–3479, 2004.

Angelini, C., Cutillo, L., De Canditiis, D., Mutarelli, M., and Pensky, M. BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, 9(1):415, 2008.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796, 2000.

Asai, S., Ohta, K., and Yoshioka, H. MAPK signaling regulates nitric oxide and NADPH oxidase-dependent oxidative bursts in Nicotiana benthamiana. *Plant Cell*, 20(5):1390–1406, 2008.

Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W.-L., Gomez-Gomez, L., Boller, T., Ausubel, F. M., and Sheen, J. MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature*, 415(6875):977–983, 2002.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9, 2000.

Asselbergh, B., De Vleesschauwer, D., and Höfte, M. Global switches and fine-tuning-ABA modulates plant pathogen defense. *Molecular Plant-Microbe Interactions*, 21(6):709–719, 2008.

Auble, D. T., Wang, D., Post, K. W., and Hahn, S. Molecular analysis of the SNF2/SWI2 protein family member MOT1, an ATP-driven enzyme that dissociates TATA-binding protein from DNA. *Molecular and Cellular Biology*, 17(8): 4842–4851, 1997.

Axelos, M., Bardet, C., Liboz, T., Le Van Thai, A., Curie, C., and Lescure, B. The gene family encoding the Arabidopsis thaliana translation elongation factor EF-1α: molecular cloning, characterization and expression. *Molecular and General Genetics MGG*, 219(1-2):106–112, 1989.

Badis, G., Chan, E. T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C. D., Gossett, A. J., Hasinoff, M. J., Warren, C. L., et al. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Molecular Cell*, 32(6):878–887, 2008.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(suppl 2):W369–W373, 2006.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

Bari, R. and Jones, J. D. G. Role of plant hormones in plant defence responses. *Plant Molecular Biology*, 69(4):473–488, 2009.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N. P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., et al. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell*, 24(10):3949–3965, 2012.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–56, 2005.

Bechtold, U., Penfold, C. A., Jenkins, D. J., Legaie, R., Moore, J. D., Lawson, T., Matthews, J. S., Vialet-Chabrand, S. R., Baxter, L., Subramaniam, S., Florance, H., Sambles, C., Salmon, D. L., Feil, R., Bowden, L., Hill, C., Baker, N. R., Lunn, J. E., Finkenstädt, B., Denby, K. J., Mead, A., Buchanan-Wollaston, V., Beynon, J., Wild, D. L., Rand, D. A., Ott, S., Smirnoff, N., and Mullineaux, P. M. Temporal analysis of physiological, metabolite and transcriptome responses during drought identifies distinct early and late phases in Arabidopsis. in preparation.

Belkhadir, Y., Yang, L., Hetzel, J., Dangl, J. L., and Chory, J. The growth–defense pivot: crisis management in plants mediated by LRR-RK surface receptors. *Trends in Biochemical Sciences*, 39(10):447–456, 2014.

Belling, K. C., Tanaka, M., Dalgaard, M. D., Nielsen, J. E., Nielsen, H. B., Brunak, S., Almstrup, K., and Leffers, H. Transcriptome profiling of mice testes following low dose irradiation. *Reproductive Biology and Endocrinology*, 11:50, 2013.

Bender, C. L., Palmer, D. A., Peñaloza-Vázquez, A., Rangaswamy, V., and Ullrich, M. Biosynthesis and regulation of coronatine, a non-host-specific phytotoxin produced by Pseudomonas syringae. In *Plant-Microbe Interactions*, pages 321–341. Springer, 1998.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Berger, M. F. and Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, 4(3):393–411, 2009.

Bergmann, S., Ihmels, J., and Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(3):031902, 2003.

Berr, A., Shafiq, S., and Shen, W.-H. Histone modifications in transcriptional activation during plant development. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1809(10):567–576, 2011.

Berr, A., Shafiq, S., Pinon, V., Dong, A., and Shen, W.-H. The trxG family histone methyltransferase SET DOMAIN GROUP 26 promotes flowering via a distinctive genetic pathway. *The Plant Journal*, 81(2):316–328, 2015.

Birkenbihl, R. P., Diezel, C., and Somssich, I. E. Arabidopsis WRKY33 is a key transcriptional regulator of hormonal and metabolic responses toward Botrytis cinerea infection. *Plant Physiology*, 159(1):266–285, 2012.

Bland, J. M. and Altman, D. G. Multiple Significance Tests: the Bonferroni Method. *BMJ: British Medical Journal*, 310(6973):170, 1995.

Bleecker, A. B. and Kende, H. Ethylene: a gaseous signal molecule in plants. *Annual Review of Cell and Developmental Biology*, 16(1):1–18, 2000.

Block, A. and Alfano, J. R. Plant targets for Pseudomonas syringae type III effectors: virulence targets or guarded decoys? *Current Opinion in Microbiology*, 14 (1):39–46, 2011.

Block, A., Li, G., Fu, Z. Q., and Alfano, J. R. Phytopathogen type III effector weaponry and their plant targets. *Current Opinion in Plant Biology*, 11(4):396–403, 2008.

Böhm, H., Albert, I., Fan, L., Reinhard, A., and Nürnberger, T. Immune receptor complexes at the plant cell surface. *Current Opinion in Plant Biology*, 20:47–54, 2014.

Boller, T. and Felix, G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology*, 60:379–406, 2009.

Boller, T. and He, S. Y. Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science*, 324 (5928):742, 2009.

Bonfig, K. B., Schreiber, U., Gabler, A., Roitsch, T., and Berger, S. Infection with virulent and avirulent P. syringae strains differentially affects photosynthesis and sink metabolism in Arabidopsis leaves. *Planta*, 225(1):1–12, 2006.

Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., and Görlach, J. Growth stage-based phenotypic analysis of Arabidopsis a model for high throughput functional genomics in plants. *Plant Cell*, 13(7): 1499–1510, 2001.

Bozkurt, T. O., Schornack, S., Banfield, M. J., and Kamoun, S. Oomycetes, effectors, and all that jazz. *Current Opinion in Plant Biology*, 15(4):483–492, 2012.

Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y. S., Penfold, C. A., Jenkins, D., Zhang, C., Morris, K., Jenner, C., Jackson, S., Thomas, B., Tabrett, A., Legaie, R., Moore, J. D., Wild, D. L., Ott, S., Rand, D., Beynon, J., Denby, K., Mead, A., and Buchanan-Wollaston, V. High-Resolution Temporal Profiling of Transcripts during Arabidopsis Leaf Senescence Reveals a Distinct Chronology of Processes and Regulation. *Plant Cell*, 23(3):873–94, 2011.

Bretz, J., Losada, L., Lisboa, K., and Hutcheson, S. W. Lon protease functions as a negative regulator of type III protein secretion in Pseudomonas syringae. *Molecular Microbiology*, 45(2):397–409, 2002.

Brodersen, P., Petersen, M., Bjørn Nielsen, H., Zhu, S., Newman, M.-A., Shokat, K. M., Rietz, S., Parker, J., and Mundy, J. Arabidopsis MAP kinase 4 regulates salicylic acid-and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4. *The Plant Journal*, 47(4):532–546, 2006.

Broekaert, W. F., Delauré, S. L., De Bolle, M. F. C., and Cammue, B. P. A. The role of ethylene in host-pathogen interactions. *Annual Review of Phytopathology*, 44:393–416, 2006.

Brooks, D. M., Bender, C. L., and Kunkel, B. N. The Pseudomonas syringae phytotoxin coronatine promotes virulence by overcoming salicylic acid-dependent defences in Arabidopsis thaliana. *Molecular Plant Pathology*, 6(6):629–639, 2005.

Brown, P., Baxter, L., Hickman, R., Beynon, J., Moore, J. D., and Ott, S. MEME-LaB: Motif Analysis in Clusters. *Bioinformatics*, 29(13):1696–7, 2013.

Brown, S. A., Weirich, C. S., Newton, E. M., and Kingston, R. E. Transcriptional activation domains stimulate initiation and elongation at different times and via different residues. *The EMBO Journal*, 17(11):3146–3154, 1998.

Bruce, W. B., Edmeades, G. O., and Barker, T. C. Molecular and physiological approaches to maize improvement for drought tolerance. *Journal of Experimental Botany*, 53(366):13–25, 2002.

Buscaill, P. and Rivas, S. Transcriptional control of plant defence responses. *Current Opinion in Plant Biology*, 20:35–46, 2014.

Busk, P. K. and Pages, M. Regulation of abscisic acid-induced transcription. *Plant Molecular Biology*, 37(3):425–435, 1998.

C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012–2018, 1998.

Causier, B. and Davies, B. Analysing protein-protein interactions with the yeast two-hybrid system. *Plant Molecular Biology*, 50(6):855–870, 2002.

Causier, B., Ashworth, M., Guo, W., and Davies, B. The TOPLESS interactome: a framework for gene repression in Arabidopsis. *Plant Physiology*, 158(1):423–438, 2012.

Chang, S. S., Park, S. K., Kim, B. C., Kang, B. J., Kim, D. U., and Nam, H. G. Stable genetic transformation of Arabidopsis thaliana by Agrobacterium inoculation in planta. *The Plant Journal*, 5(4):551–558, 1994.

Chen, H., Zhang, D., Guo, J., Wu, H., Jin, M., Lu, Q., Lu, C., and Zhang, L. A Psb27 homologue in Arabidopsis thaliana is required for efficient repair of photodamaged photosystem II. *Plant Molecular Biology*, 61(4-5):567–75, 2006.

Chen, L.-Q., Hou, B.-H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X.-Q., Guo, W.-J., Kim, J.-G., Underwood, W., Chaudhuri, B., et al. Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature*, 468(7323):527–532, 2010.

Chen, L.-Q., Qu, X.-Q., Hou, B.-H., Sosso, D., Osorio, S., Fernie, A. R., and Frommer, W. B. Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science*, 335(6065):207–211, 2012a.

Chen, M. Q., Zhang, A. H., Zhang, Q., Zhang, B. C., Nan, J., Li, X., Liu, N., Qu, H., Lu, C. M., Sudmorgen, Zhou, Y. H., Xu, Z. H., and Bai, S. N. Arabidopsis NMD3 is required for nuclear export of 60S ribosomal subunits and affects secondary cell wall thickening. *PLoS One*, 7(4):e35904, 2012b.

Cheng, Y. and Church, G. M. Biclustering of expression data. In *ISMB*, volume 8, pages 93–103, 2000.

Chini, A., Fonseca, S., Chico, J. M., Fernández-Calvo, P., and Solano, R. The ZIM domain mediates homo-and heteromeric interactions between Arabidopsis JAZ proteins. *The Plant Journal*, 59(1):77–87, 2009.

Chinnusamy, V., Gong, Z., and Zhu, J.-K. Abscisic Acid-mediated Epigenetic Processes in Plant Development and Stress Responses. *Journal of Integrative Plant Biology*, 50(10):1187–1195, 2008.

Chisholm, S. T., Dahlbeck, D., Krishnamurthy, N., Day, B., Sjolander, K., and Staskawicz, B. J. Molecular characterization of proteolytic cleavage sites of the Pseudomonas syringae effector AvrRpt2. *Proceedings of the National Academy of Sciences*, 102(6):2087–2092, 2005.

Clay, N. K., Adio, A. M., Denoux, C., Jander, G., and Ausubel, F. M. Glucosinolate metabolites required for an Arabidopsis innate immune response. *Science*, 323 (5910):95–101, 2009.

Clough, S. J. and Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The Plant Journal*, 16(6):735–743, 1998.

Coaker, G., Falick, A., and Staskawicz, B. Activation of a phytopathogenic bacterial effector protein by a eukaryotic cyclophilin. *Science*, 308(5721):548–550, 2005.

Collier, S. M. and Moffett, P. NB-LRRs work a bait and switch on pathogens. *Trends in Plant Science*, 14(10):521–529, 2009.

Colmenares, A. J., Aleu, J., Duran-Patron, R., Collado, I. G., and Hernandez-Galan, R. The putative role of botrydial and related metabolites in the infection mechanism of Botrytis cinerea. *Journal of Chemical Ecology*, 28(5):997–1005, 2002.

Cramer, P., Bushnell, D. A., and Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science*, 292(5523):1863–1876, 2001.

Crick, F. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Cubas, P., Lauter, N., Doebley, J., and Coen, E. The TCP Domain: a Motif Found in Proteins Regulating Plant Growth and Development. *Plant Journal*, 18(2): 215–22, 1999.

Cui, H., Wang, Y., Xue, L., Chu, J., Yan, C., Fu, J., Chen, M., Innes, R. W., and Zhou, J.-M. Pseudomonas syringae effector protein AvrB perturbs Arabidopsis hormone signaling by activating MAP kinase 4. *Cell Host & Microbe*, 7(2):164–175, 2010.

Cui, H., Tsuda, K., and Parker, J. E. Effector-triggered immunity: from pathogen perception to robust defense. *Annual Review of Plant Biology*, 66:487–511, 2015.

Cunnac, S., Lindeberg, M., and Collmer, A. Pseudomonas syringae type III secretion system effectors: repertoires in search of functions. *Current Opinion in Microbiology*, 12(1):53–60, 2009.

Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R., and Abrams, S. R. Abscisic acid: emergence of a core signaling network. *Annual Reviews of Plant Biology*, 61:651–679, 2010.

Dangl, J. L., Horvath, D. M., and Staskawicz, B. J. Pivoting the plant immune system from dissection to deployment. *Science*, 341(6147):746–751, 2013.

Dantuma, N. P., Groothuis, T. A., Salomons, F. A., and Neefjes, J. A dynamic ubiquitin equilibrium couples proteasomal activity to chromatin remodeling. *The Journal of Cell Biology*, 173(1):19–26, 2006.

de Pater, S., Greco, V., Pham, K., Memelink, J., and Kijne, J. Characterization of a Zinc-Dependent Transcriptional Activator from Arabidopsis. *Nucleic Acids Research*, 24(23):4624–31, 1996.

de Torres, M., Mansfield, J. W., Grabov, N., Brown, I. R., Ammouneh, H., Tsiamis, G., Forsyth, A., Robatzek, S., Grant, M., and Boch, J. Pseudomonas syringae effector AvrPtoB suppresses basal defence in Arabidopsis. *The Plant Journal*, 47 (3):368–382, 2006.

de Torres-Zabala, M., Littlejohn, G., Jayaraman, S., Studholme, D., Bailey, T., Lawson, T., Tillich, M., Licht, D., Bölter, B., Delfino, L., Truman, W., Mansfield, J., Smirnoff, N., , and Grant, M. Photosynthesis is central to plant defence and pathogen effectors target the chloroplast. *Nature Plants*, in press.

de Torres-Zabala, M., Truman, W., Bennett, M. H., Lafforgue, G., Mansfield, J. W., Rodriguez Egea, P., Bögre, L., and Grant, M. Pseudomonas syringae pv. tomato hijacks the Arabidopsis abscisic acid signalling pathway to cause disease. *The EMBO Journal*, 26(5):1434–1443, 2007.

de Torres-Zabala, M., Bennett, M. H., Truman, W. H., and Grant, M. R. Antagonism between salicylic and abscisic acid reflects early host–pathogen conflict and moulds plant defence responses. *The Plant Journal*, 59(3):375–386, 2009.

De Vos, M., Van Zaanen, W., Koornneef, A., Korzelius, J. P., Dicke, M., Van Loon, L. C., and Pieterse, C. M. J. Herbivore-induced resistance against microbial pathogens in Arabidopsis. *Plant Physiology*, 142(1):352–363, 2006.

de Waard, M. A., Andrade, A. C., Hayashi, K., Schoonbeek, H.-j., Stergiopoulos, I., and Zwiers, L.-H. Impact of fungal drug transporters on fungicide sensitivity, multidrug resistance and virulence. *Pest Management Science*, 62(3):195–207, 2006.

Delker, C., Stenzel, I., Hause, B., Miersch, O., Feussner, I., and Wasternack, C. Jasmonate Biosynthesis in Arabidopsis thaliana — Enzymes, Products, Regulation. *Plant Biology*, 8(3):297–306, 2006.

Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A. M., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J. S., Hope, I. A., Tissenbaum, H. A., Mango, S. E., and Walhout, A. J. A Gene-Centered C. elegans Protein-DNA Interaction Network. *Cell*, 125(6):1193–205, 2006.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A. J. M. A gateway-compatible yeast one-hybrid system. *Genome Research*, 14(10b):2093–2101, 2004.

Deslandes, L., Olivier, J., Peeters, N., Feng, D. X., Khounlotham, M., Boucher, C., Somssich, I., Genin, S., and Marco, Y. Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. *Proceedings of the National Academy of Sciences*, 100(13):8024–8029, 2003.

Després, C., DeLong, C., Glaze, S., Liu, E., and Fobert, P. R. The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. *Plant Cell*, 12(2):279–290, 2000.

Dolfini, D., Gatta, R., and Mantovani, R. NF-Y and the transcriptional activation of CCAAT promoters. *Critical Reviews in Biochemistry and Molecular Biology*, 47(1):29–49, 2012.

Dong, X., Jiang, Z., Peng, Y.-L., and Zhang, Z. Revealing Shared and Distinct Gene Network Organization in Arabidopsis Immune Responses by Integrative Analysis. *Plant Physiology*, 167(3):1186–1203, 2015.

Droby, S. and Lichter, A. Post-harvest Botrytis infection: etiology, development and management. In *Botrytis: biology, pathology and control*, pages 349–367. Springer, 2007.

Du, J., Johnson, L. M., Groth, M., Feng, S., Hale, C. J., Li, S., Vashisht, A. A., Gallego-Bartolome, J., Wohlschlegel, J. A., Patel, D. J., et al. Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Molecular Cell*, 55 (3):495–504, 2014.

Dudler, R. Manipulation of host proteasomes as a virulence mechanism of plant pathogens. *Annual Review of Phytopathology*, 51(1):521, 2013.

Durrant, W. E. and Dong, X. Systemic acquired resistance. *Annual Review of Phytopathology*, 42:185–209, 2004.

Dutton, M. V. and Evans, C. S. Oxalate production by fungi: its role in pathogenicity and ecology in the soil environment. *Canadian Journal of Microbiology*, 42(9): 881–895, 1996.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, Ü. V. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292, 2013.

Eulgem, T., Rushton, P. J., Robatzek, S., and Somssich, I. E. The WRKY Superfamily of Plant Transcription Factors. *Trends in Plant Science*, 5(5):199–206, 2000.

Eulgem, T. and Somssich, I. E. Networks of WRKY transcription factors in defense signaling. *Current Opinion in Plant Biology*, 10(4):366–371, 2007.

Feng, F. and Zhou, J.-M. Plant–bacterial pathogen interactions mediated by type III effectors. *Current Opinion in Plant Biology*, 15(4):469–476, 2012.

Feuillet, C. and Keller, B. High gene density is conserved at syntenic loci of small and large grass genomes. *Proceedings of the National Academy of Sciences*, 96 (14):8265–8270, 1999.

Flavell, R. Role of model plant species. In *Plant Genomics*, pages 1–18. Springer, 2009.

Fliegmann, J., Mithöfer, A., Wanner, G., and Ebel, J. An ancient enzyme domain hidden in the putative β-glucan elicitor receptor of soybean may play an active part in the perception of pathogen-associated molecular patterns during broad host resistance. *Journal of Biological Chemistry*, 279(2):1132–1140, 2004.

Franco-Zorrilla, J. M., Løpez-Vidriero, I., Carrasco, J. L., Godoy, M., Vera, P., and Solano, R. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, 111(6):2367–2372, 2014.

Fujita, M., Fujita, Y., Maruyama, K., Seki, M., Hiratsu, K., Ohme-Takagi, M., Tran, L.-S. P., Yamaguchi-Shinozaki, K., and Shinozaki, K. A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *The Plant Journal*, 39(6):863–876, 2004.

Fujita, Y., Yoshida, T., and Yamaguchi-Shinozaki, K. Pivotal role of the AREB/ABF-SnRK2 pathway in ABRE-mediated transcription in response to osmotic stress in plants. *Physiologia Plantarum*, 147(1):15–27, 2013.

Garcia, M. E., Lynch, T., Peeters, J., Snowden, C., and Finkelstein, R. A small plant-specific protein family of ABI five binding proteins (AFPs) regulates stress response in germinating Arabidopsis seeds and seedlings. *Plant Molecular Biology*, 67(6):643–58, 2008.

Gibbs, H. K., Ruesch, A. S., Achard, F., Clayton, M. K., Holmgren, P., Ramankutty, N., and Foley, J. A. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proceedings of the National Academy of Sciences*, 107(38):16732–16737, 2010.

Glazebrook, J. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology*, 43:205–227, 2005.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., and Johnston, M. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

Göhre, V., Spallek, T., Häweker, H., Mersmann, S., Mentzel, T., Boller, T., de Torres, M., Mansfield, J. W., and Robatzek, S. Plant pattern-recognition receptor FLS2 is directed for degradation by the bacterial ubiquitin ligase AvrPtoB. *Current Biology*, 18(23):1824–1832, 2008.

Gomez-Cadenas, A., Arbona, V., Jacas, J., Primo-Millo, E., and Talon, M. Abscisic acid reduces leaf abscission and increases salt tolerance in citrus plants. *Journal of Plant Growth Regulation*, 21(3):234–240, 2002.

Gømez-Gømez, L. and Boller, T. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Molecular Cell*, 5(6): 1003–1011, 2000.

Gømez-Gømez, L., Felix, G., and Boller, T. A single locus determines sensitivity to bacterial flagellin in Arabidopsis thaliana. *The Plant Journal*, 18(3):277–284, 1999.

Gophna, U., Ron, E. Z., and Graur, D. Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312:151–163, 2003.

Govrin, E. M. and Levine, A. The hypersensitive response facilitates plant infection by the necrotrophic pathogen Botrytis cinerea. *Current Biology*, 10(13):751–757, 2000.

Grant, C. E., Bailey, T. L., and Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

Grant, M., Brown, I., Adams, S., Knight, M., Ainslie, A., and Mansfield, J. The RPM1 plant disease resistance gene facilitates a rapid and sustained increase in cytosolic calcium that is necessary for the oxidative burst and hypersensitive cell death. *The Plant Journal*, 23(4):441–450, 2000.

Grant, M. R., Kazan, K., and Manners, J. M. Exploiting pathogens' tricks of the trade for engineering of plant disease resistance: challenges and opportunities. *Microbial Biotechnology*, 6(3):212–222, 2013.

Greeff, C., Roux, M., Mundy, J., and Petersen, M. Receptor-like kinase complexes in plant innate immunity. *Frontiers in Plant Science*, 3, 2012.

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One*, 5(10):e13397–e13397, 2010.

Groll, M., Schellenberg, B., Bachmann, A. S., Archer, C. R., Huber, R., Powell, T. K., Lindow, S., Kaiser, M., and Dudler, R. A plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism. *Nature*, 452(7188): 755–758, 2008.

Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature*, 389(6649):349–352, 1997.

Guo, H. and Ecker, J. R. Plant responses to ethylene gas are mediated by SCF EBF1/EBF2-dependent proteolysis of EIN3 transcription factor. *Cell*, 115(6): 667–677, 2003.

Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural & Molecular Biology*, 11(5):394, 2004.

Han, S.-K., Sang, Y., Rodrigues, A., Wu, M.-F., Rodriguez, P. L., Wagner, D., et al. The SWI2/SNF2 chromatin remodeling ATPase BRAHMA represses abscisic acid responses in the absence of the stress stimulus in Arabidopsis. *The Plant Cell*, 24 (12):4892–4906, 2012.

Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., and Dimopoulos, G. Bayesian Coclustering of Anopheles Gene Expression Time Series: Study of Immune Defense Response to Multiple Experimental Challenges. *Proceedings of the National Academy of Sciences*, 102(47):16939–44, 2005.

Heard, N. A., Holmes, C. C., and Stephens, D. A. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, 101(473):18–29, 2006.

Henfling, J., Bostock, R., and Kuc, J. Effect of abscisic acid on rishitin and lubimin accumulation and resistance to Phytophthora infestans and Cladosporium cucumerinum in potato tuber tissue slices. *Phytopathology*, 70(11):1074–1078, 1980.

Herms, D. A. and Mattson, W. J. The dilemma of plants: to grow or defend. *Quarterly Review of Biology*, pages 283–335, 1992.

Heyndrickx, K. S. and Vandepoele, K. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiology*, 159(3):884–901, 2012.

Heyndrickx, K. S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. *Plant Cell*, 26(10):3894–3910, 2014.

Hickman, R., Hill, C., Penfold, C. A., Breeze, E., Bowden, L., Moore, J. D., Zhang, P., Jackson, A., Cooke, E., Bewicke-Copley, F., et al. A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves. *The Plant Journal*, 75(1):26–39, 2013.

Higo, K., Ugawa, Y., Iwamoto, M., and Higo, H. PLACE: a Database of Plant Cis-Acting Regulatory DNA Elements. *Nucleic Acids Research*, 26(1):358–9, 1998.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, 27(1):297–300, 1999.

Hirano, S. S. and Upper, C. D. Population biology and epidemiology of Pseudomonas syringae. *Annual Review of Phytopathology*, 28(1):155–177, 1990.

Hobo, T., Kowyama, Y., and Hattori, T. A bZIP factor, TRAB1, interacts with VP1 and mediates abscisic acid-induced transcription. *Proceedings of the National Academy of Sciences*, 96(26):15348–15353, 1999.

Hochheimer, A. and Tjian, R. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes & Development*, 17(11):1309–1320, 2003.

Hofius, D., Schultz-Larsen, T., Joensen, J., Tsitsigiannis, D. I., Petersen, N. H. T., Mattsson, O., Jrgensen, L. B., Jones, J. D. G., Mundy, J., and Petersen, M. Autophagic components contribute to hypersensitive cell death in Arabidopsis. *Cell*, 137(4):773–783, 2009.

Howard, B. E., Hu, Q., Babaoglu, A. C., Chandra, M., Borghi, M., Tan, X., He, L., Winter-Sederoff, H., Gassmann, W., Veronese, P., et al. High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. *PloS One*, 8(10):e74183, 2013.

Huang, W., Pérez-García, P., Pokhilko, A., Millar, A. J., Antoshechkin, I., Riechmann, J. L., and Mas, P. Mapping the Core of the Arabidopsis Circadian Clock Defines the Network Structure of the Oscillator. *Science*, 336(6077):75–79, 2012.

Ifuku, K., Endo, T., Shikanai, T., and Aro, E. M. Structure of the chloroplast NADH dehydrogenase-like complex: nomenclature for nuclear-encoded subunits. *Plant and Cell Physiology*, 52(9):1560–8, 2011.

Ifuku, K., Yamamoto, Y., Ono, T.-a., Ishihara, S., and Sato, F. PsbP protein, but not PsbQ protein, is essential for the regulation and stabilization of photosystem II in higher plants. *Plant Physiology*, 139(3):1175–1184, 2005.

Ikeda, K., Steger, D. J., Eberharter, A., and Workman, J. L. Activation domain-specific and general transcription stimulation by native histone acetyltransferase complexes. *Molecular and Cellular Biology*, 19(1):855–863, 1999.

Innes, R. W., Bent, A. F., Kunkel, B. N., Bisgrove, S. R., and Staskawicz, B. Molecular analysis of avirulence gene avrRpt2 and identification of a putative regulatory sequence common to all known Pseudomonas syringae avirulence genes. *Journal of Bacteriology*, 175(15):4859–4869, 1993.

Ito, S., Song, Y. H., Josephson-Day, A. R., Miller, R. J., Breton, G., Olmstead, R. G., and Imaizumi, T. FLOWERING BHLH transcriptional activators control expression of the photoperiodic flowering regulator CONSTANS in Arabidopsis. *Proceedings of the National Academy of Sciences*, 109(9):3582–3587, 2012.

Iyer-Pascuzzi, A. S. and McCouch, S. R. Recessive resistance genes and the Oryza sativa-Xanthomonas oryzae pv. oryzae pathosystem. *Molecular Plant-Microbe Interactions*, 20(7):731–739, 2007.

Jackson, M. B. and Osborne, D. J. Ethylene, the natural regulator of leaf abscission. *Nature*, 225(5237):1019–22, 1970.

Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Jain, A. K. and Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

Jensen, M. K., Hagedorn, P. H., Torres-Zabala, D., Grant, M. R., Rung, J. H., Collinge, D. B., Lyngkjaer, M. F., et al. Transcriptional regulation by an NAC (NAM–ATAF1, 2–CUC2) transcription factor attenuates ABA signalling for efficient basal defence towards Blumeria graminis f. sp. hordei in Arabidopsis. *The Plant Journal*, 56(6):867–880, 2008.

Jin, G., Davey, M. C., Ertl, J. R., Chen, R., Yu, Z.-t., Daniel, S. G., Becker, W. M., and Chen, C.-m. Interaction of DNA-binding proteins with the 5-flanking region of a cytokinin-responsive cucumber hydroxypyruvate reductase gene. *Plant Molecular Biology*, 38(5):713–723, 1998.

Jin, Q., Thilmony, R., Zwiesler-Vollick, J., and He, S.-Y. Type III protein secretion in Pseudomonas syringae. *Microbes and Infection*, 5(4):301–310, 2003.

Jones, A. M., Bennett, M. H., Mansfield, J. W., and Grant, M. Analysis of the defence phosphoproteome of Arabidopsis thaliana using differential mass tagging. *Proteomics*, 6(14):4155–4165, 2006a.

Jones, A. M., Thomas, V., Bennett, M. H., Mansfield, J., and Grant, M. Modifications to the Arabidopsis defense proteome occur prior to significant transcriptional change in response to inoculation with Pseudomonas syringae. *Plant Physiology*, 142(4):1603–1620, 2006b.

Jones, J. D. G. and Dangl, J. L. The plant immune system. *Nature*, 444(7117): 323–329, 2006.

Journot-Catalino, N., Somssich, I. E., Roby, D., and Kroj, T. The transcription factors WRKY11 and WRKY17 act as negative regulators of basal resistance in Arabidopsis thaliana. *The Plant Cell*, 18(11):3289–3302, 2006.

Ju, C., Yoon, G. M., Shemansky, J. M., Lin, D. Y., Ying, Z. I., Chang, J., Garrett, W. M., Kessenbrock, M., Groth, G., and Tucker, M. L. CTR1 phosphorylates the central regulator EIN2 to control ethylene hormone signaling from the ER membrane to the nucleus in Arabidopsis. *Proceedings of the National Academy of Sciences*, 109(47):19486–19491, 2012.

Kadota, Y., Sklenar, J., Derbyshire, P., Stransfeld, L., Asai, S., Ntoukakis, V., Jones, J. D., Shirasu, K., Menke, F., Jones, A., et al. Direct regulation of the NADPH oxidase RBOHD by the PRR-associated kinase BIK1 during plant immunity. *Molecular Cell*, 54(1):43–55, 2014.

Kagale, S. and Rozwadowski, K. EAR motif-mediated transcriptional repression in plants: an underlying mechanism for epigenetic regulation of gene expression. *Epigenetics*, 6(2):141–146, 2011.

Kaku, H., Nishizawa, Y., Ishii-Minami, N., Akimoto-Tomiyama, C., Dohmae, N., Takio, K., Minami, E., and Shibuya, N. Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor. *Proceedings of the National Academy of Sciences*, 103(29):11086–11091, 2006.

Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., Riddle, N. C., Verbsky, M. L., and Richards, E. J. Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics*, 163(3):1109–1122, 2003.

Katagiri, F., Thilmony, R., and He, S. Y. The Arabidopsis thaliana-Pseudomonas syringae interaction. *The Arabidopsis Book*, 1, 2002.

Kazan, K. Negative regulation of defence and stress genes by EAR-motif-containing repressors. *Trends in Plant Science*, 11(3):109–112, 2006.

Kazan, K. and Lyons, R. Intervention of phytohormone pathways by pathogen effectors. *The Plant Cell*, 26(6):2285–2309, 2014.

Kazan, K. and Manners, J. M. MYC2: the master in action. *Molecular Plant*, 6(3): 686–703, 2013.

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579, 2003.

Keren, N., Ohkawa, H., Welsh, E. A., Liberton, M., and Pakrasi, H. B. Psb29, a conserved 22-kD protein, functions in the biogenesis of photosystem II complexes in Synechocystis and Arabidopsis. *The Plant Cell*, 17(10):2768–2781, 2005.

Kesarwani, M., Yoo, J., and Dong, X. Genetic interactions of TGA transcription factors in the regulation of pathogenesis-related genes and disease resistance in Arabidopsis. *Plant Physiology*, 144(1):336–346, 2007.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. The AtGenExpress Global Stress Expression Data Set: Protocols, Evaluation and Model Data Analysis of UV-B Light, Drought and Cold Stress Responses. *Plant Journal*, 50(2):347–63, 2007.

Kim, Y., Han, S., Choi, S., and Hwang, D. Inference of Dynamic Networks Using Time-Course Data. *Briefings in Bioinformatics*, 2013. 10.1093/bib/bbt028.

King, E. O., Ward, M. K., and Raney, D. E. Two simple media for the demonstration of pyocyanin and fluorescin. *The Journal of Laboratory and Clinical Medicine*, 44 (2):301–307, 1954.

Kleinboelting, N., Huep, G., Kloetgen, A., Viehoever, P., and Weisshaar, B. GABI-Kat SimpleSearch: new features of the Arabidopsis thaliana T-DNA mutant database. *Nucleic Acids Research*, page gkr1047, 2011.

Klemm, S. L. Causal structure identification in nonlinear dynamical systems. *Department of Engineering, University of Cambridge, UK*, 2008.

Kornberg, R. D. Eukaryotic transcriptional control. *Trends in Biochemical Sciences*, 24(12):M46–M49, 1999.

Korves, T. M. and Bergelson, J. A developmental response to pathogen infection in Arabidopsis. *Plant Physiology*, 133(1):339–347, 2003.

Kreps, J. A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., and Harper, J. F. Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, 130(4):2129–2141, 2002.

Kubori, T., Matsushima, Y., Nakamura, D., Uralil, J., Lara-Tejero, M., Sukhan, A., Galán, J. E., and Aizawa, S.-I. Supramolecular structure of the Salmonella typhimurium type III protein secretion system. *Science*, 280(5363):602–605, 1998.

Laurie-Berry, N., Joardar, V., Street, I. H., and Kunkel, B. N. The Arabidopsis thaliana JASMONATE INSENSITIVE 1 gene is required for suppression of salicylic acid-dependent defenses during infection by Pseudomonas syringae. *Molecular Plant-Microbe Interactions*, 19(7):789–800, 2006.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology*, 28(2):149–156, 2010.

Lee, M. W., Jelenska, J., and Greenberg, J. T. Arabidopsis proteins important for modulating defense responses to Pseudomonas syringae that secrete HopW1-1. *Plant Journal*, 54(3):452–65, 2008.

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., Shim, J. E., Shim, H., Kim, H., and Kim, C. AraNet v2: an improved database of co-functional gene networks for the study of Afrabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Research*, page gku1053, 2014.

Leon-Reyes, A., Spoel, S. H., De Lange, E. S., Abe, H., Kobayashi, M., Tsuda, S., Millenaar, F. F., Welschen, R. A. M., Ritsema, T., and Pieterse, C. M. J. Ethylene modulates the role of NONEXPRESSOR OF PATHOGENESIS-RELATED GENES1 in cross talk between salicylate and jasmonate signaling. *Plant Physiology*, 149(4):1797–1809, 2009.

Lewis, J. D., Lee, A. H.-Y., Hassan, J. A., Wan, J., Hurley, B., Jhingree, J. R., Wang, P. W., Lo, T., Youn, J.-Y., and Guttman, D. S. The Arabidopsis ZED1 pseudokinase is required for ZAR1-mediated immunity induced by the Pseudomonas syringae type III effector HopZ1a. *Proceedings of the National Academy of Sciences*, 110(46):18722–18727, 2013.

Lewis, L. A., Polanski, K., de Torres-Zabala, M., Jayaraman, S., Bowden, L., Moore, J., Penfold, C. A., Jenkins, D. J., Hill, C., Baxter, L., Kulasekaran, S., Truman, W., Littlejohn, G., Prusinska, J., Mead, A., Steinbrenner, J., Hickman, R., Rand, D., Wild, D. L., Ott, S., Buchanan-Wollaston, V., Smirnoff, N., Denby, K., Beynon, J., and Grant, M. Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in Arabidopsis leaves following infection with Pseudomonas syringae pv. tomato DC3000. *Plant Cell*, in press.

Li, B., Jiang, S., Yu, X., Cheng, C., Chen, S., Cheng, Y., Yuan, J. S., Jiang, D., He, P., and Shan, L. Phosphorylation of trihelix transcriptional repressor ASR3 by MAP KINASE4 negatively regulates Arabidopsis immunity. *The Plant Cell*, 27(3):839–856, 2015.

Li, J., Brader, G., and Palva, E. T. The WRKY70 transcription factor: a node

of convergence for jasmonate-mediated and salicylate-mediated signals in plant defense. *Plant Cell*, 16(2):319–331, 2004.

Li, J., Brader, G., Kariola, T., and Tapio Palva, E. WRKY70 modulates the selection of signaling pathways in plant defense. *The Plant Journal*, 46(3):477–491, 2006.

Li, X., Ye, Y., Wu, Q., and Ng, M. K. Multifactv: Finding modules from higher-order gene expression profiles with time dimension. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

Lindeberg, M., Cunnac, S., and Collmer, A. Pseudomonas syringae type III effector repertoires: last words in endless arguments. *Trends in Microbiology*, 20(4):199–208, 2012.

Lindgren, P. B., Peet, R. C., and Panopoulos, N. J. Gene cluster of Pseudomonas syringae pv." phaseolicola" controls pathogenicity of bean plants and hypersensitivity of nonhost plants. *Journal of Bacteriology*, 168(2):512–522, 1986.

Liu, J., Elmore, J. M., Fuglsang, A. T., Palmgren, M. G., Staskawicz, B. J., and Coaker, G. RIN4 functions with plasma membrane H+-ATPases to regulate stomatal apertures during pathogen attack. *PLoS Biology*, 7(6):1400, 2009.

Liu, X., Chen, C.-Y., Wang, K.-C., Luo, M., Tai, R., Yuan, L., Zhao, M., Yang, S., Tian, G., Cui, Y., et al. PHYTOCHROME INTERACTING FACTOR3 associates with the histone deacetylase HDA15 in repression of chlorophyll biosynthesis and photosynthesis in etiolated Arabidopsis seedlings. *The Plant Cell*, 25 (4):1258–1273, 2013.

Lopez, J. A., Sun, Y., Blair, P. B., and Mukhtar, M. S. TCP three-way handshake: linking developmental processes with plant immunity. *Trends in Plant Science*, 20(4):238–245, 2015.

Lorenzo, O., Piqueras, R., Sánchez-Serrano, J. J., and Solano, R. ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. *Plant Cell*, 15(1):165–178, 2003.

Lorenzo, O., Chico, J. M., Sánchez-Serrano, J. J., and Solano, R. JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in Arabidopsis. *Plant Cell*, 16(7):1938–1950, 2004.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biology*, 1(1):1–37, 2000.

Ma, J. and Ptashne, M. The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. *Cell*, 50(1):137–142, 1987.

Ma, K.-W., Flores, C., and Ma, W. Chromatin configuration as a battlefield in plant-bacteria interactions. *Plant Physiology*, 157(2):535–543, 2011.

Ma, S., Gong, Q., and Bohnert, H. J. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research*, 17(11):1614–1625, 2007.

Macho, A. P. and Zipfel, C. Plant PRRs and the activation of innate immune signaling. *Molecular Cell*, 54(2):263–272, 2014.

Macho, A. P. and Zipfel, C. Targeting of plant pattern recognition receptor-triggered immunity by bacterial type-III secretion system effectors. *Current Opinion in Microbiology*, 23:14–22, 2015.

Mackey, D., Holt, B. F., Wiig, A., and Dangl, J. L. RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis. *Cell*, 108(6):743–754, 2002.

Mackey, D., Belkhadir, Y., Alonso, J. M., Ecker, J. R., and Dangl, J. L. Arabidopsis RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell*, 112(3):379–389, 2003.

Madeira, S. C., Teixeira, M. C., Sa-Correia, I., and Oliveira, A. L. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 7(1):153–165, 2010.

Maere, S., Heymans, K., and Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9, 2005.

Maere, S., Van Dijck, P., and Kuiper, M. Extracting Expression Modules from Perturbational Gene Expression Compendia. *BMC Systems Biology*, 2:33, 2008.

Mansfield, J. W. From bacterial avirulence genes to effector functions via the hrp delivery system: an overview of 25 years of progress in our understanding of plant innate immunity. *Molecular Plant Pathology*, 10(6):721–734, 2009.

Martín-Trillo, M. and Cubas, P. TCP genes: a family snapshot ten years later. *Trends in Plant Science*, 15(1):31–39, 2010.

Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., Goda, H., Shimada, Y., Yoshida, S., Shinozaki, K., and Yamaguchi-Shinozaki, K. Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. *The Plant Journal*, 38(6):982–993, 2004.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(suppl 1):D108–D110, 2006.

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D., and Koornneef, M. Arabidopsis thaliana: a model plant for genome analysis. *Science*, 282(5389):662–682, 1998.

Melotto, M., Underwood, W., Koczan, J., Nomura, K., and He, S. Y. Plant stomata function in innate immunity against bacterial invasion. *Cell*, 126(5):969–980, 2006.

Meng, J., Gao, S.-J., and Huang, Y. Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics*, 25(12):1521–1527, 2009.

Menges, M., Dóczi, R., Ökrész, L., Morandini, P., Mizzi, L., Soloviev, M., Murray, J. A. H., and Bögre, L. Comprehensive gene expression atlas for the Arabidopsis MAP kinase signalling pathways. *New Phytologist*, 179(3):643–662, 2008.

Mengiste, T. Plant immunity to necrotrophs. *Annual Review of Phytopathology*, 50: 267–294, 2012.

Mengiste, T., Chen, X., Salmeron, J., and Dietrich, R. The BOTRYTIS SUSCEPTIBLE1 gene encodes an R2R3MYB transcription factor protein that is required for biotic and abiotic stress responses in Arabidopsis. *Plant Cell*, 15(11):2551–2565, 2003.

Miranda, O. R., You, C.-C., Phillips, R., Kim, I.-B., Ghosh, P. S., Bunz, U. H. F., and Rotello, V. M. Array-based sensing of proteins using conjugated polymers. *Journal of the American Chemical Society*, 129(32):9856–9857, 2007.

Mitchell, K., Brown, I., Knox, P., and Mansfield, J. The role of cell wall-based defences in the early restriction of non-pathogenic hrp mutant bacteria in Arabidopsis. *Phytochemistry*, 112:139–150, 2015.

Mithöfer, A., Ebel, J., and Felle, H. H. Cation fluxes cause plasma membrane depolarization involved in β-glucan elicitor-signaling in soybean roots. *Molecular Plant-Microbe Interactions*, 18(9):983–990, 2005.

Mittler, R. Abiotic stress, the field environment and stress combination. *Trends in Plant Science*, 11(1):15–19, 2006.

Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H., and Shibuya, N. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. *Proceedings of the National Academy of Sciences*, 104(49):19613–19618, 2007.

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffiere, A., Yan, S., Dominguez, H., and Thompson, B. M. The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *The ISME Journal*, 2(3):321–334, 2008.

Mou, Z., Fan, W., and Dong, X. Inducers of plant systemic acquired resistance regulate NPR1 function through redox changes. *Cell*, 113(7):935–944, 2003.

Mukhtar, M. S., Carvunis, A.-R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M. T., et al. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, 333(6042):596–601, 2011.

Murakami, R., Ifuku, K., Takabayashi, A., Shikanai, T., Endo, T., and Sato, F. Functional dissection of two Arabidopsis PsbO proteins: PsbO1 and PsbO2. *FEBS Journal*, 272(9):2165–75, 2005.

Murray, S. L., Ingle, R. A., Petersen, L. N., and Denby, K. J. Basal resistance against Pseudomonas syringae in Arabidopsis involves WRKY53 and a protein with homology to a nematode resistance protein. *Molecular Plant-Microbe Interactions*, 20(11):1431–1438, 2007.

Näär, A. M., Lemon, B. D., and Tjian, R. Transcriptional coactivator complexes. *Annual Review of Biochemistry*, 70(1):475–501, 2001.

Nambara, E. and Marion-Poll, A. Abscisic acid biosynthesis and catabolism. *Annual Review of Plant Biology*, 56:165–185, 2005.

Onouchi, H., Igeño, M. I., Périlleux, C., Graves, K., and Coupland, G. Mutagenesis of plants overexpressing CONSTANS demonstrates novel interactions among Arabidopsis flowering-time genes. *Plant Cell*, 12(6):885–900, 2000.

Orphanides, G., Lagrange, T., and Reinberg, D. The general transcription factors of RNA polymerase II. *Genes & Development*, 10(21):2657, 1996.

Ouwerkerk, P. B. and Meijer, A. H. Yeast one-hybrid screening for DNA-protein interactions. *Current Protocols in Molecular Biology*, Chapter 12:Unit 12 12, 2001.

Palusa, S. G., Ali, G. S., and Reddy, A. S. N. Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *The Plant Journal*, 49(6):1091–1107, 2007.

Pandey, S. P. and Somssich, I. E. The role of WRKY transcription factors in plant immunity. *Plant Physiology*, 150(4):1648–1655, 2009.

Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

Patel, S. and Dinesh-Kumar, S. P. Arabidopsis ATG6 is required to limit the pathogen-associated cell death response. *Autophagy*, 4(1):20–27, 2008.

Pauwels, L. and Goossens, A. Fine-tuning of early events in the jasmonate response. *Plant Signaling & Behavior*, 3(10):846–847, 2008.

Pauwels, L., Barbero, G. F., Geerinck, J., Tilleman, S., Grunewald, W., Pérez, A. C., Chico, J. M., Bossche, R. V., Sewell, J., Gil, E., et al. NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature*, 464(7289):788–791, 2010.

Penfold, C. A. and Wild, D. L. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

Penfold, C. A., Shifaz, A., Brown, P. E., Nicholson, A., and Wild, D. L. CSI: a non-parametric Bayesian approach to network inference from multiple perturbed time series gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 14(3):307–310, 2015.

Penfold, C. A., Hill, C., McHattie, S., Breeze, E., Windram, O., Kim, Y.-s., Kiddle, S., Cooke, E., Jackson, A., Prusinska, J., Moore, J., Jenkins, D., Finkenstadt, B., Ott, S., Rand, D., Beynon, J., Wild, D., Buchanan-Wollaston, V., and Denby, K. J. Network-directed screening identifies key regulators of the Arabidopsis response to pathogen infection and developmental senescence. in preparation.

Peng, Y. and Jahroudi, N. The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. *Journal of Biological Chemistry*, 278(10):8385–8394, 2003.

Penninckx, I. A., Eggermont, K., Terras, F. R., Thomma, B. P., De Samblanx, G. W., Buchala, A., Métraux, J.-P., Manners, J. M., and Broekaert, W. F. Pathogen-induced systemic activation of a plant defensin gene in Arabidopsis follows a salicylic acid-independent pathway. *Plant Cell*, 8(12):2309–2323, 1996.

Penninckx, I. A. M. A., Thomma, B. P. H. J., Buchala, A., Métraux, J.-P., and Broekaert, W. F. Concomitant activation of jasmonate and ethylene response pathways is required for induction of a plant defensin gene in Arabidopsis. *Plant Cell*, 10(12):2103–2113, 1998.

Pieterse, C. M. J., Leon-Reyes, A., Van der Ent, S., and Van Wees, S. C. M. Networking by small-molecule hormones in plant immunity. *Nature Chemical Biology*, 5(5):308–316, 2009.

Pieterse, C. M., Van der Does, D., Zamioudis, C., Leon-Reyes, A., and Van Wees, S. C. Hormonal modulation of plant immunity. *Annual Review of Cell and Developmental Biology*, 28:489–521, 2012.

Pokhilko, A., Hodge, S. K., Stratford, K., Knox, K., Edwards, K. D., Thomson, A. W., Mizuno, T., and Millar, A. J. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, 6(1):416, 2010.

Pokhilko, A., Fernández, A. P., Edwards, K. D., Southern, M. M., Halliday, K. J., and Millar, A. J. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology*, 8(1):574, 2012.

Polanski, K., Rhodes, J., Hill, C., Zhang, P., Jenkins, D. J., Kiddle, S. J., Jironkin, A., Beynon, J., Buchanan-Wollaston, V., Ott, S., and Denby, K. J. Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics*, 2014.

Pré, M., Atallah, M., Champion, A., De Vos, M., Pieterse, C. M. J., and Memelink, J. The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiology*, 147(3):1347–1357, 2008.

Pruneda-Paz, J. L., Breton, G., Nagel, D. H., Kang, S. E., Bonaldi, K., Doherty, C. J., Ravelo, S., Galli, M., Ecker, J. R., and Kay, S. A. A genome-scale resource

for the functional characterization of Arabidopsis transcription factors. *Cell Reports*, 8(2):622–632, 2014.

Rabbani, M. A., Maruyama, K., Abe, H., Khan, M. A., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. *Plant Physiology*, 133(4):1755–1767, 2003.

Rahme, L. G., Mindrinos, M. N., and Panopoulos, N. J. Plant and environmental sensory signals control the expression of hrp genes in Pseudomonas syringae pv. phaseolicola. *Journal of Bacteriology*, 174(11):3499–3507, 1992.

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22(1), 2008.

Rasmussen, P. E., Goulding, K. W. T., Brown, J. R., Grace, P. R., Janzen, H. H., and Körschens, M. Long-term agroecosystem experiments: Assessing agricultural sustainability and global change. *Science*, 282(5390):893–896, 1998.

Redman, J. C., Haas, B. J., Tanimoto, G., and Town, C. D. Development and Evaluation of an Arabidopsis Whole Genome Affymetrix Probe Array. *Plant Journal*, 38(3):545–61, 2004.

Reiss, D. J., Baliga, N. S., and Bonneau, R. Integrated Biclustering of Heterogeneous Genome-Wide Datasets for the Inference of Global Regulatory Networks. *BMC Bioinformatics*, 7:280, 2006.

Rhee, S. Y. and Mutwil, M. Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19(4):212–221, 2014.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., and Montoya, M. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, 31(1):224–228, 2003.

Rhodes, J. *Identifying gene regulatory networks common to multiple plant stress responses.* PhD thesis, University of Warwick, 2012.

Ricardi, M. M., González, R. M., Zhong, S., Domínguez, P. G., Duffy, T., Turjanski, P. G., Salter, J. D. S., Alleva, K., Carrari, F., and Giovannoni, J. J. Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor. *BMC Plant Biology*, 14(1):29, 2014.

Riechmann, J., Heard, J., Martin, G., Reuber, L., Jiang, C.-Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O., Samaha, R., et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–2110, 2000.

Rizhsky, L., Liang, H., and Mittler, R. The combined effect of drought stress and heat shock on gene expression in tobacco. *Plant Physiology*, 130(3):1143–1151, 2002.

Robatzek, S., Chinchilla, D., and Boller, T. Ligand-induced endocytosis of the pattern recognition receptor FLS2 in Arabidopsis. *Genes & Development*, 20(5): 537–542, 2006.

Robert-Seilaniantz, A., Grant, M., and Jones, J. D. Hormone Crosstalk in Plant Disease and Defense: More than Just Jasmonate-Salicylate Antagonism. *Annual Review of Phytopathology*, 49:317–43, 2011.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing. *Nature Methods*, 4(8):651–7, 2007.

Roine, E., Wei, W., Yuan, J., Nurmiaho-Lassila, E.-L., Kalkkinen, N., Romantschuk, M., and He, S. Y. Hrp pilus: an hrp-dependent bacterial surface appendage produced by Pseudomonas syringae pv. tomato DC3000. *Proceedings of the National Academy of Sciences*, 94(7):3459–3464, 1997.

Ron, M. and Avni, A. The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. *Plant Cell*, 16(6):1604–1615, 2004.

Rosebrock, T. R., Zeng, L., Brady, J. J., Abramovitch, R. B., Xiao, F., and Martin, G. B. A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. *Nature*, 448(7151):370–374, 2007.

Rosslenbroich, H.-J. and Stuebler, D. Botrytis cinereahistory of chemical control and novel fungicides for its management. *Crop Protection*, 19(8):557–561, 2000.

Roush, R. and Tabashnik, B. E. *Pesticide resistance in arthropods*. Springer Science & Business Media, 2012.

Saez, A., Rodrigues, A., Santiago, J., Rubio, S., and Rodriguez, P. L. HAB1–SWI3B interaction reveals a link between abscisic acid signaling and putative SWI/SNF chromatin-remodeling complexes in Arabidopsis. *The Plant Cell*, 20(11):2972–2988, 2008.

Sakamoto, H., Maruyama, K., Sakuma, Y., Meshi, T., Iwabuchi, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiology*, 136(1):2734–2746, 2004.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, 2004.

Sarnowski, T. J., Świez, S., Pawlikowska, K., Kaczanowski, S., Jerzmanowski, A., et al. AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucleic Acids Research*, 30(15):3412–3421, 2002.

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J., and Wild, D. L. R/BHC: Fast Bayesian Hierarchical Clustering for Microarray Data. *BMC Bioinformatics*, 10:242, 2009.

Schouten, A., Tenberge, K. B., Vermeer, J., Stewart, J., Wagemakers, L., Williamson, B., and Van Kan, J. A. L. Functional analysis of an extracellular catalase of Botrytis cinerea. *Molecular Plant Pathology*, 3(4):227–238, 2002.

Schulze, A. and Downward, J. Navigating gene expression using microarraysa technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

Sclep, G., Allemeersch, J., Liechti, R., De Meyer, B., Beynon, J., Bhalerao, R., Moreau, Y., Nietfeld, W., Renou, J. P., Reymond, P., Kuiper, M. T. R., and Hilson, P. CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC Bioinformatics*, 8, 2007.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., Dietrich, B., Ho, P., Bacwaden, J., and Ko, C. A high-throughput Arabidopsis reverse genetics system. *Plant Cell*, 14(12):2985–2994, 2002.

Shan, L., He, P., Li, J., Heese, A., Peck, S. C., Nrnberger, T., Martin, G. B., and Sheen, J. Bacterial effectors target the common signaling partner BAK1 to disrupt multiple MAMP receptor-signaling complexes and impede plant immunity. *Cell Host & Microbe*, 4(1):17–27, 2008.

Sheard, L. B., Tan, X., Mao, H., Withers, J., Ben-Nissan, G., Hinds, T. R., Kobayashi, Y., Hsu, F.-F., Sharon, M., and Browse, J. Jasmonate perception by inositol-phosphate-potentiated COI1-JAZ co-receptor. *Nature*, 468(7322):400–405, 2010.

Shinozaki, K., Yamaguchi-Shinozaki, K., and Seki, M. Regulatory network of gene expression in the drought and cold stress responses. *Current Opinion in Plant Biology*, 6(5):410–417, 2003.

Siebert, S., Döll, P., Hoogeveen, J., Faures, J. M., Frenken, K., and Feick, S. Development and validation of the global map of irrigation areas. *Hydrology and Earth System Sciences Discussions Discussions*, 2(4):1299–1327, 2005.

Smyth, G. K., Michaud, J., and Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

Spanu, P., Grosskopf, D. G., Felix, G., and Boller, T. The apparent turnover of 1-aminocyclopropane-1-carboxylate synthase in tomato cells is regulated by protein phosphorylation and dephosphorylation. *Plant Physiology*, 106(2):529–535, 1994.

Spoel, S. H., Koornneef, A., Claessens, S. M. C., Korzelius, J. P., Van Pelt, J. A., Mueller, M. J., Buchala, A. J., Métraux, J.-P., Brown, R., and Kazan, K. NPR1 modulates cross-talk between salicylate-and jasmonate-dependent defense pathways through a novel function in the cytosol. *Plant Cell*, 15(3):760–770, 2003.

Sridhar, V. V., Kapoor, A., Zhang, K., Zhu, J., Zhou, T., Hasegawa, P. M., Bressan, R. A., and Zhu, J.-K. Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. *Nature*, 447(7145):735–738, 2007.

Staats, M., van Baarlen, P., Schouten, A., van Kan, J. A. L., and Bakker, F. T. Positive selection in phytotoxic protein-encoding genes of Botrytis species. *Fungal Genetics and Biology*, 44(1):52–63, 2007.

Stael, S., Kmiecik, P., Willems, P., Van Der Kelen, K., Coll, N. S., Teige, M., and Van Breusegem, F. Plant innate immunity–sunny side up? *Trends in Plant Science*, 20(1):3–11, 2015.

Staswick, P. E. and Tiryaki, I. The oxylipin signal jasmonic acid is activated by an enzyme that conjugates it to isoleucine in Arabidopsis. *Plant Cell*, 16(8): 2117–2127, 2004.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. A Robust Bayesian Two-Sample Test for Detecting Intervals of Differential Gene Expression in Microarray Time Series. *Journal of Computational Biology*, 17(3):355–67, 2010.

Supper, J., Strauch, M., Wanke, D., Harter, K., and Zell, A. EDISA: Extracting Biclusters from Multiple Time-Series of Gene Expression Profiles. *BMC Bioinformatics*, 8:334, 2007.

Taoka, K.-i., Kaya, H., Nakayama, T., Araki, T., Meshi, T., and Iwabuchi, M. Identification of three kinds of mutually related composite elements conferring S phase-specific transcriptional activation. *The Plant Journal*, 18(6):611–623, 1999.

Tchagang, A. B., Pan, Y., Famili, F., Tewfik, A. H., and Benos, P. V. Biclustering of dna microarray data: Theory, evaluation, and applications. 2010.

Thilmony, R., Underwood, W., and He, S. Y. Genome-wide transcriptional analysis of the Arabidopsis thaliana interaction with the plant pathogen Pseudomonas syringae pv. tomato DC3000 and the human pathogen Escherichia coli O157: H7. *The Plant Journal*, 46(1):34–53, 2006.

Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Toh, H. and Horimoto, K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, 18(2):287–297, 2002.

Tran, L.-S. P., Nakashima, K., Sakuma, Y., Simpson, S. D., Fujita, Y., Maruyama, K., Fujita, M., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *The Plant Cell*, 16(9):2481–2498, 2004.

Trujillo, M., Ichimura, K., Casais, C., and Shirasu, K. Negative Regulation of PAMP-Triggered Immunity by an E3 Ubiquitin Ligase Triplet in Arabidopsis. *Current Biology*, 18(18):1396–1401, 2008.

Truman, W., Zabala, M. T., and Grant, M. Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance. *The Plant Journal*, 46(1):14–33, 2006.

Tsuda, K. and Katagiri, F. Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. *Current Opinion in Plant Biology*, 13 (4):459–465, 2010.

Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., and Katagiri, F. Network properties of robust immunity in plants. *PLoS Genetics*, 5(12):e1000772, 2009.

Tukey, J. W. Exploratory data analysis. 1977.

Tuteja, N. Abscisic acid and abiotic stress signaling. *Plant Signaling & Behavior*, 2 (3):135–138, 2007.

Ullah, H., Chen, J.-G., Young, J. C., Im, K.-H., Sussman, M. R., and Jones, A. M. Modulation of cell proliferation by heterotrimeric G protein in Arabidopsis. *Science*, 292(5524):2066–2069, 2001.

Umezawa, T., Sugiyama, N., Mizoguchi, M., Hayashi, S., Myouga, F., Yamaguchi-Shinozaki, K., Ishihama, Y., Hirayama, T., and Shinozaki, K. Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in Arabidopsis. *Proceedings of the National Academy of Sciences*, 106(41):17588–17593, 2009.

United Nations. World Population Prospects - The 2012 Revision, 2012.

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K., and Yamaguchi-Shinozaki, K. Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proceedings of the National Academy of Sciences*, 97(21): 11632–11637, 2000.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N. J. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment*, 32(12):1633–1651, 2009.

van Kan, J. A. L. Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends in Plant Science*, 11(5):247–253, 2006.

Van Loon, L. C., Rep, M., and Pieterse, C. M. J. Significance of inducible defense-related proteins in infected plants. *Annual Review of Phytopathology*, 44:135–162, 2006.

Vincent, A. C. and Struhl, K. ACR1, a yeast ATF/CREB repressor. *Molecular and Cellular Biology*, 12(12):5394–5405, 1992.

Vlot, A. C., Dempsey, D. A., and Klessig, D. F. Salicylic acid, a multifaceted hormone to combat disease. *Annu Rev Phytopathol*, 47:177–206, 2009.

Wan, J., Zhang, X.-C., Neece, D., Ramonell, K. M., Clough, S., Kim, S.-y., Stacey, M. G., and Stacey, G. A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in Arabidopsis. *The Plant Cell*, 20(2):471–481, 2008.

Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

Ward, J. L., Forcat, S., Beckmann, M., Bennett, M., Miller, S. J., Baker, J. M., Hawkins, N. D., Vermeer, C. P., Lu, C., Lin, W., et al. The metabolic transition during disease following infection of Arabidopsis thaliana by Pseudomonas syringae pv. tomato. *The Plant Journal*, 63(3):443–457, 2010.

Weinstock-Guttman, B., Badgett, D., Patrick, K., Hartrich, L., Santos, R., Hall, D., Baier, M., Feichter, J., and Ramanathan, M. Genomic Effects of IFN-Beta in Multiple Sclerosis Patients. *Journal of Immunology*, 171(5):2694–702, 2003.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., and Cook, K. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6): 1431–1443, 2014.

Welchen, E. and Gonzalez, D. H. Overrepresentation of Elements Recognized by TCP-Domain Transcription Factors in the Upstream Regions of Nuclear Genes Encoding Components of the Mitochondrial Oxidative Phosphorylation Machinery. *Plant Physiology*, 141(2):540–5, 2006.

Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., McDonald, N., Wiley, K., Bader, K. C., Gläßer, C., et al. Convergent targeting of a common

host protein-network by pathogen effectors from three kingdoms of life. *Cell Host & Microbe*, 16(3):364–375, 2014.

Wheeler, T. and von Braun, J. Climate change impacts on global food security. *Science*, 341(6145):508–513, 2013.

Williamson, B., Tudzynski, B., Tudzynski, P., and van Kan, J. A. L. Botrytis cinerea: the cause of grey mould disease. *Molecular Plant Pathology*, 8(5):561–580, 2007.

Win, J., Chaparro-Garcia, A., Belhaj, K., Saunders, D., Yoshida, K., Dong, S., Schornack, S., Zipfel, C., Robatzek, S., Hogenhout, S., et al. Effector biology of plant-associated organisms: concepts and perspectives. In *Cold Spring Harbor symposia on quantitative biology*, volume 77, pages 235–247. Cold Spring Harbor Laboratory Press, 2012.

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., Kiddle, S. J., Rhodes, J., Atwell, S., Kliebenstein, D. J., Kim, Y. S., Stegle, O., Borgwardt, K., Zhang, C., Tabrett, A., Legaie, R., Moore, J., Finkenstadt, B., Wild, D. L., Mead, A., Rand, D., Beynon, J., Ott, S., Buchanan-Wollaston, V., and Denby, K. J. Arabidopsis Defense against Botrytis cinerea: Chronology and Regulation Deciphered by High-Resolution Temporal Transcriptomic Analysis. *Plant Cell*, 24(9):3530–57, 2012.

Windram, O., Penfold, C. A., and Denby, K. J. Network modeling to understand plant immunity. *Annual Review of Phytopathology*, 52:93–111, 2014.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. TRANSFAC: a Database on Transcription Factors and Their DNA Binding Sites. *Nucleic Acids Research*, 24 (1):238–41, 1996.

Wittenberg, G., Levitan, A., Klein, T., Dangoor, I., Keren, N., and Danon, A. Knockdown of the Arabidopsis thaliana chloroplast protein disulfide isomerase 6 results in reduced levels of photoinhibition and increased D1 synthesis in high light. *The Plant Journal*, 78(6):1003–1013, 2014.

Wu, H., Kerr, M. K., Cui, X., and Churchill, G. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. In *The Analysis of Gene Expression Data*, Statistics for Biology and Health, chapter 14, pages 313–341. Springer New York, 2003.

Wu, L., Chen, H., Curtis, C., and Fu, Z. Q. Go in for the kill: How plants deploy effector-triggered immunity to combat pathogens. *Virulence*, 5(7):710–721, 2014.

Xie, C., Zhou, X., Deng, X., and Guo, Y. PKS5, a SNF1-related kinase, interacts with and phosphorylates NPR1, and modulates expression of WRKY38 and WRKY62. *Journal of Genetics and Genomics*, 37(6):359–369, 2010.

Xin, X.-F. and He, S. Y. Pseudomonas syringae pv. tomato DC3000: a model pathogen for probing disease susceptibility and hormone signaling in plants. *Annual Review of Phytopathology*, 51:473–498, 2013.

Xu, L., Thali, M., and Schaffner, W. Upstream box/TATA box order is the major determinant of the direction of transcription. *Nucleic Acids Res*, 19(24):6699–6704, 1991.

Xu, X., Chen, C., Fan, B., and Chen, Z. Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors. *The Plant Cell Online*, 18(5):1310–1326, 2006.

Yamamoto, S., Nakano, T., Suzuki, K., and Shinshi, H. Elicitor-induced activation of transcription via W box-related cis-acting elements from a basic chitinase gene by WRKY transcription factors in tobacco. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1679(3):279–287, 2004.

Yang, J., Penfold, C. A., Grant, M. R., and Rattray, M. Inferring the perturbation time from biological time course data. *Bioinformatics*, in press.

Yasuda, M., Ishikawa, A., Jikumaru, Y., Seki, M., Umezawa, T., Asami, T., Maruyama-Nakashita, A., Kudo, T., Shinozaki, K., and Yoshida, S. Antagonistic interaction between systemic acquired resistance and the abscisic acid-mediated abiotic stress response in Arabidopsis. *Plant Cell*, 20(6):1678–1692, 2008.

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biology*, 5 (7):R48, 2004.

Yona, G., Dirks, W., Rahman, S., and Lin, D. M. Effective similarity measures for expression profiles. *Bioinformatics*, 22(13):1616–22, 2006.

Yuen, C. Y., Matsumoto, K. O., and Christopher, D. A. Variation in the Subcellular Localization and Protein Folding Activity among Arabidopsis thaliana Homologs of Protein Disulfide Isomerase. *Biomolecules*, 3(4):848–869, 2013.

Zentner, G. E. and Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nature Structural & Molecular Biology*, 20(3):259–266, 2013.

Zhang, J., Shao, F., Li, Y., Cui, H., Chen, L., Li, H., Zou, Y., Long, C., Lan, L., and Chai, J. A Pseudomonas syringae effector inactivates MAPKs to suppress PAMP-induced immunity in plants. *Cell Host & Microbe*, 1(3):175–185, 2007.

Zhang, S., Wang, K., Ashby, C., Chen, B., and Huang, X. *A Unified Adaptive Co-Identification Framework for High-D Expression Data*, volume 7632 of *Lecture Notes in Computer Science*, chapter 6, pages 59–70. Springer Berlin Heidelberg, 2012.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W. L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., and Jacobsen, S. E. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, 126(6):1189–1201, 2006.

Zhang, X., Han, X., Shi, R., Yang, G., Qi, L., Wang, R., and Li, G. Arabidopsis cysteine-rich receptor-like kinase 45 positively regulates disease resistance to Pseudomonas syringae. *Plant Physiology and Biochemistry*, 73:383–391, 2013.

Zhao, C., Avci, U., Grant, E. H., Haigler, C. H., and Beers, E. P. XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem. *The Plant Journal*, 53(3): 425–436, 2008.

Zhao, Y., Brickner, J. R., Majid, M. C., and Mosammaparast, N. Crosstalk between ubiquitin and other post-translational modifications on chromatin during double-strand break repair. *Trends in Cell Biology*, 24(7):426–434, 2014.

Zheng, X.-y., Spivey, N. W., Zeng, W., Liu, P.-P., Fu, Z. Q., Klessig, D. F., He, S. Y., and Dong, X. Coronatine promotes Pseudomonas syringae virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation. *Cell Host & Microbe*, 11(6):587–596, 2012.

Zheng, Z.-L. and Zhao, Y. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to 'Candidatus Liberibacter asiaticus' infection. *BMC Genomics*, 14(1):27, 2013.

Zhu, J.-K. Genetic analysis of plant salt tolerance using Arabidopsis. *Plant Physiology*, 124(3):941–948, 2000.

Zhu, Q., Zhang, J., Gao, X., Tong, J., Xiao, L., Li, W., and Zhang, H. The Arabidopsis AP2/ERF transcription factor RAP2.6 participates in ABA, salt and osmotic stress responses. *Gene*, 457(1):1–12, 2010.

Zhu, Z., An, F., Feng, Y., Li, P., Xue, L., Mu, A., Jiang, Z., Kim, J.-M., To, T. K., and Li, W. Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. *Proceedings of the National Academy of Sciences*, 108(30):12539–12544, 2011.

Zipfel, C. Plant pattern-recognition receptors. *Trends in immunology*, 35(7):345–351, 2014.

Zipfel, C. and Felix, G. Plants and animals: a different taste for microbes? *Current Opinion in Plant Biology*, 8(4):353–360, 2005.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E. J., Jones, J. D. G., Felix, G., and Boller, T. Bacterial disease resistance in Arabidopsis through flagellin perception. *Nature*, 428(6984):764–767, 2004.

Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J. D. G., Boller, T., and Felix, G. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. *Cell*, 125(4):749–760, 2006.